AN ANALYSIS OF TECHNIQUES FOR CONTROLLING FAMILYWISE ERROR IN THE CONTEXT OF A PARTIAL NULL HYPOTHESIS

Thesis

Submitted to

The School of Arts and Sciences of the

UNIVERSITY OF DAYTON

in Partial Fulfillment of the Requirements for

The Degree

Master of Arts in Psychology

by

Robert Charles Dix

UNIVERSITY OF DAYTON

Dayton, Ohio

July, 2003

APPROVED BY:

CONCURRENCE

ABSTRACT

AN ANALYSIS OF TECHNIQUES FOR CONTROLLING FAMILYWISE ERROR IN THE CONTEXT OF A PARTIAL NULL HYPOTHESIS

Name: Dix, Robert Charles
University of Dayton, 2003

Advisor: Dr. David Biers

The purpose of the current study was to examine 10 post hoc techniques for controlling familywise error rate in tests of simple comparisons for factorial designs. Familywise error under both the complete and partial null hypothesis, plus Type II error, were investigated for varying factorial designs. The study used 3X3, 3X5, and 5X3 designs in a Monte Carlo study and manipulated sample size, pattern of variability of the coefficients within each true simple effect, and effect size of the interaction. The results suggest that the approach used should depend on whether or not there is a true simple effect. Aside from using Tukey Overall on every occasion, the results indicate that when there is no true simple effect Bonferroni should be used, and when there is a true simple effect either Tukey Row, Bonferroni Row, or Dual Bonferroni should be the techniques of choice. This division makes sense because when there is no true simple effect, paying a penalty at the level of simple effects leads to testing no simple comparisons within the null simple effect. Alternatively, when there is a true simple effect and the error rate penalty is not sufficient, all simple comparisons will be tested without any protection unless a simple comparisons penalty is applied. When viewing this information from a

practical viewpoint it becomes apparent that the information is not very useful. The recommendation, therefore, would be to use the Keppel approach for the 3 x 3 designs when the means for the true simple effects are not at the extremes. For any other situation tested in this study, however, Tukey overall was the only technique which consistently controlled familywise error. Tukey overall is not without a cost however, as there is a large loss in power to detect significance. Overall, this study confirms Keppel's conclusion that effects are so complex that it is difficult to reach a conclusion when a researcher exceeds three levels per variable. Therefore when planning a factorial design it is best to limit the number of levels to three per variable.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

CHAPTER I

INTRODUCTION

The analysis of variance (ANOVA) is the most widely used statistical procedure by psychologists. However, one of the shortcomings of the F-test is that it does not determine the locus of significance when there are more than two conditions. It is for this reason most researchers perform some sort of post-hoc analysis following a significant F-test. There is a wide variety of post-hoc analytic strategies which can be employed in analysis of the data. Studies (Jaccard, Becker & Wood, 1984; Keselman, Keselman, & Games, 1991) have thoroughly studied the Type I error, Type II error, familywise error, and Power of these procedures within the context of single factor designs. However, as Keppel (1991) states, there is a paucity of research in use of these techniques with factorial designs (particularly in post-hoc testing of the significance of the interaction).

Biers and his colleagues (Anthony 1995; Brake 1994; Reising 1993) have examined Type I, Type II, and familywise error rates for post-hoc analytic procedures within the context of a 3 x 3, 3 x 5 and 5 x 3 factorial design. The present study seeks to reanalyze existing data for additional post hoc procedures and for inclusion of Type II error associated with the 5 x 3 study.

Overview

A common post-hoc analytic approach associated with analysis of variance in a two factor design consists of three successive analyses (omnibus, simple effects, simple comparisons), each dependent on the previous test being significant. This approach is recommended by Keppel (1991) and is termed the "filter" approach by Biers and his colleagues.

Perhaps the easiest way to understand this approach is to look at an example; Figure 1 is a graphical representation of a 3x3 factorial design in which there are three levels of variable A (A1, A2, A3) crossed with three levels of variable B (B1, B2, B3) to form nine conditions (A1B1, A2B1....A3B3).

Variable	**A1**	**A2**	**A3**	Main Effect of Variable B
B1	A1B1	A2B1	A3B1	
B2	A2B2	A2B2	A3B2	
B3	A1B3	A2B3	A3B3	
Main Effect of Variable A				

Figure 1. A 3x3 factorial Design

A 3 x 3 factorial ANOVA would be performed in the above design in which the main effects (A alone and B alone) and the interaction effect (A x B) are first tested for significance. If a main effect is found to be significant, then the researcher knows that variables 'A' or 'B' (or both) has an independent effect on the dependent variable. If the interaction effect is significant, the experimenter knows that the effect of one independent variable (e.g. A) is different across levels of the other independent variable (e.g. B). If there are more than two levels to the IV, then it is not possible to determine where the

differences lie without proceeding to the analysis of simple effects (Rosnow & Rosenthal, 1989), the first stage of post-hoc analyses. Analysis of simple effects consists of determining the effects of A at a particular level of variable B. For example, variable 'A' (A1 vs. A2 vs. A3) may be studied at each row of the above table (B1, B2, and B3). In terms of the omnibus analysis, the sum of variance of the simple effect contains the variance of the main effect and of the interaction effect (Kirk, 1982). Thus:

$$SS_{A@Bj} = \Sigma(SS_A + SS_{A \times B})$$

Suppose, for the moment, that a simple effect, A @ B1 was determined to be significant. This would mean that the entire variable of A has an effect at the level of B1. The problem is if there are more than two levels of A, then it is not possible to determine which levels of A are differing at B1. In order to make that conclusion the researcher must perform the computations for the second stage of post-hoc analyses, the analysis of simple comparisons. In this stage, pairwise comparisons of the treatment means are made. Following the above example of the significant simple effect (A @ B1), and assuming only pairwise comparisons are made, the following simple comparisons would be made: A1 vs. A2 @B1, A1 vs. A3 @B1, and A2 vs. A3 @B1.

In summary, a post-hoc analysis of interaction begins with the omnibus *F* being tested. If the omnibus *F* is found to be significant the researcher would conduct an analysis of simple effects. This indicates whether or not a certain level of one independent variable differs at a specific level of another independent variable. The analysis of simple comparisons is the last step and only is performed if a significant simple effect is found. This last stage usually consists of a set of pairwise comparisons

that compares all levels of one independent variable at a given level of the other. This post-hoc analytic procedure is analogous to passing the data through a series of successive filters. A given pairwise difference is significant if and only if it makes it through all three filters. Figure 2 demonstrates the correct order for post-hoc analyses.

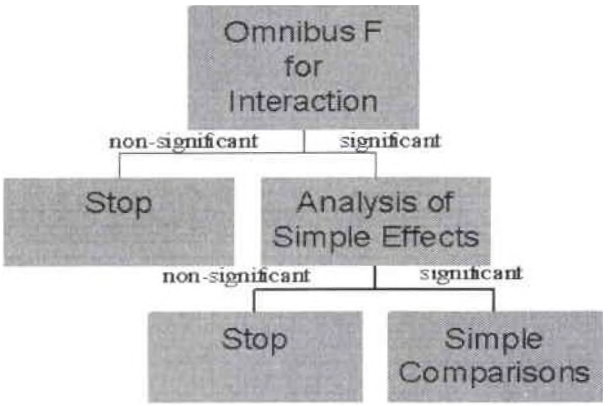

Figure 2. Steps in the Contingency Analysis Process (Adopted from Brake, 1994).

Decision Errors and Power

Basic Definitions

When the researcher tests an effect for significance, he or she concludes that the results are due to sampling error, or that there is a true treatment effect. To determine significance in an analysis of variance, the obtained *F* value is compared to the critical *F* value. During the evaluation of significance two possible errors may occur; Type I (α) and Type II (β). A Type I error is the probability of rejecting H_0 when in fact the H_0 is

true. Alternatively, a Type I error occurs when the results are attributed to the treatment effect, when the results are actually due to sampling error. Type II error occurs when the results are attributed to sampling error although there is actually a treatment effect: the probability of finding H_0 true, when in fact it is false.

Power is a concept that is associated with Type II error. Power is the probability of finding a treatment effect statistically significant, when the treatment is actually the reason for the effect. Power indicates strength of the experiment to find significant results when the treatment is accountable for the change. Power is represented as $1-\beta$; therefore as Type II (β) error increases, power decreases and vice versa.

Figure 3 demonstrates the principles of Type I error, Type II error, and power. The more stringent significance level (shown in Figure 3B) lowers the chance of a Type I error occurring. Unfortunately, this also increases the probability that any true effect will be missed (Type II error). Power = $1-\beta$, (where β is Type II error) as β increases power decreases. Reducing the probability of a Type I error in a factorial ANOVA by controlling α_{FW} (see p. 7) could decrease the sensitivity of the experiment. Taking the two previous statements into consideration, the key is to balance Types I and II error in order to achieve the optimal level of power.

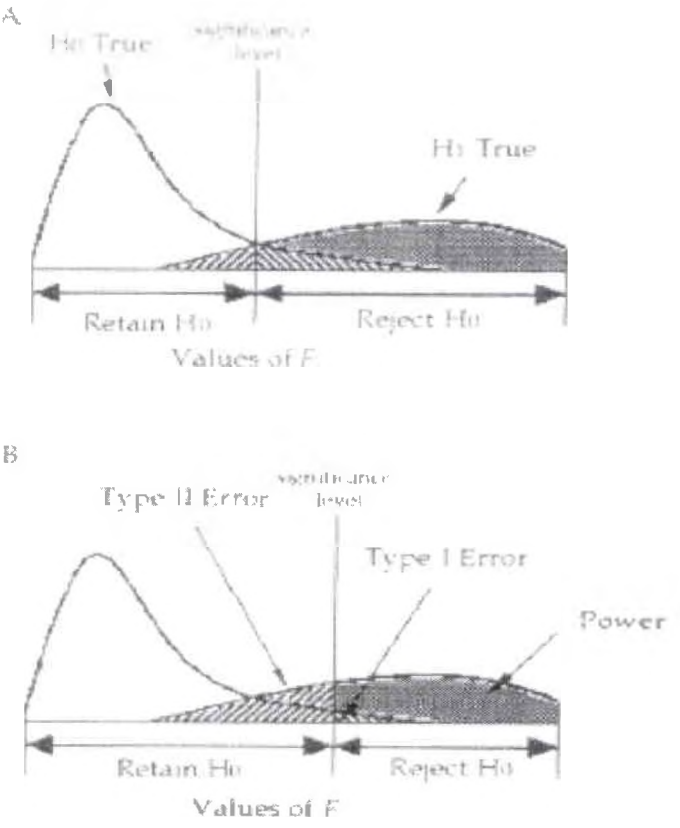

Figure 3. The Relationship Among Type I Error, Type II Error, and Power (adapted from Keppel, 1991; Brake 1994).

The Problem of Compounding Type I Error

The significance level which is adopted determines the likelihood of a Type I error in a single test of significance. A .05 significance level means that 5% of the time a Type I error will occur in the given test of significance: This is known as error rate per comparison (α_{PC}). Type I error is controlled by adopting a more stringent significance level (e.g. α=.01). Figure 3 represents this relationship.

When large numbers of statistical tests are being conducted (as in post-hoc analyses) the chance of committing at least one Type I error increases as the number of tests increases. For the 3 x 3 ANOVA previously mentioned, if the omnibus *F* were found to be significant and all three simple effects (A@B1, A@B2, and A@B3) were also significant, the next step would be to perform simple comparisons. This would mean that three simple comparisons would be conducted for each level of B; a total of nine tests. Assuming that the comparisons are orthogonal and a .05 significance level is used, the probability of making at least one Type I error in this family (α_{FW}) can be computed using the following formula:

$$\alpha_{FW} = 1-(1-\alpha_{PC})^{c}$$

where: c = the number of statistical tests

In the example,

$$\alpha_{FW} = 1-(1-.05)^{9} = .3698$$

The following formula may be used to approximate:

$$\alpha_{FW} \cong \alpha_{PC}(c)$$

From the example,

$$\alpha_{FW} \cong .05(9) \cong .4500$$

Whichever value is used, it is apparent that the probability of a Type I error occurring is larger than the acceptable .05 level. In a 3 x 4 or a 4 x 4 factorial ANOVA the values would be even greater ($1-(1-.05)^{12}$ and $1-(1-.05)^{16}$, respectively).

A common way of controlling for the compounding of familywise error is to adopt a more stringent significance level for each statistical test conducted. The Bonferroni technique is one such way to control for familywise error, defined by the formula:

$$\alpha_{PC_{ADJ}} = \alpha_{FW}/c$$

where:

$\alpha_{PC_{ADJ}}$ is each test's adjusted probability of making a Type I error.

The α_{FW} is held constant (usually .05) by the Bonferroni technique, and then adjusts the significance level for the number of tests. From the example:

$$\alpha_{PC} = .05/9 = .0056$$

The above example indicates that each statistical test should be conducted at the .0056 level to maintain α_{FW} at .05. Thus, we are paying a penalty for conducting multiple tests by requiring a more stringent significance level. This penalty is termed the post-hoc error rate penalty.

The Complete Versus the Partial-Null Hypothesis

The probability of making a decision error and the power of a statistical test are of the utmost importance to any researcher. Type I error (α_{PC}) and α_{FW} assume that one is testing the complete null hypothesis. It assumes there is no treatment effect and all conditions have identical population means. The complete null can be represented as:

$$H_0 : \mu_1 = \mu_2 = \mu_3 = \mu_n$$

With the complete H_o all pairwise differences are simultaneously equal to 0.

The partial-null hypothesis is an important variation of the null hypothesis that is not often considered (Ryan 1980). The partial null hypothesis occurs when there are a number of groups that have identical means, and one or more has a different mean. This may be seen below:

$$H_0 : \mu_1 = \mu_2 = \mu_3 \neq \mu_4$$

Under the partial H_0, when one examines the pairwise differences, some are different and others are not. That is, we have non-significant effects within the context of significant differences. What happens to Type I error and familywise error within the context of these significant effects? This is an issue which has not been addressed in previous research by Biers and his colleagues.

Factors Affecting Type II Error and Power

There are three statistical elements which have an effect on Type II error and power. They are: 1) the magnitude of the treatment effect, 2) sample size, and 3) error variance. The formula for power and the associated statistic of ϕ demonstrate the effects of the three elements, as seen below:

$$\phi = \sqrt{n\frac{\sum(\mu_s - \mu_s)^2 / a}{\sigma_s^2 / a}}$$

where:

n is the sample size,

μ_I are the population treatment means,

μ_t is the mean of treatment means,

a is the number of treatment means, and

σ_s^2/a is the mean variance of the treatment populations.

The treatment magnitude is represented by the numerator in the above equation. This means that ϕ increases as the numerator increases. Thus, the likelihood of a Type II error decreases with increases in treatment magnitude and power. Another element which affects ϕ is sample size, represented by n. Like with the numerator, ϕ increases as n

increases; therefore power increases and Type II error decreases with any increase in n. Lastly, error variance is represented by the denominator. As the denominator increases the value of ϕ decreases which means power decreases and the probability of Type II error increases.

Given the above relationships, it is possible to manipulate the probability of a Type II error and thus power by changing the magnitude of the treatment effect, sample size, or error variance. Thus, in the second study, treatment magnitude and sample size will be varied to influence Type II error and power.

Methods for Controlling Familywise Error Rate

There are many techniques which may be used to perform post hoc analyses. However, it would not be practical to sample all of the techniques, so this study will look at the seven approaches which sample a broad range of techniques for controlling compounding familywise error. Generally, these techniques will fall onto one of the following three classes: 1) pay no penalty, 2) make a correction at the level of simple effects, 3) make a correction at the level of simple comparisons. This section will explain the seven techniques for controlling familywise error. Table 1 gives a summary of these techniques.

Table 1. Summary of Post-hoc Control Techniques

Post-hoc Analysis Procedure	Omnibus *F* Test of A x B Interaction	Analysis of Simple Effects	Analysis of Simple Comparisons
No Penalty			
Fisher LSD (FISH)	.05	-	.05
Keppel - Filter (KEP)	.05	.05	.05
Penalty at Simple Effects			
Bonferroni (BON)	.05	.05/3 = .017*	.05
Modified Bonferroni (MB)	.05	.10/3 = .033*	.05
Modified Bonferroni - Both (MBB)	.05	.15/6 = .025*	.05
Penalty at Simple Comparisons			
Tukey - Overall (TOvl)	.05	-	.00205*
Tukey - Row (TRow)	.05	.05	.01926*

* assuming a 3 x 3 design.

Paying No Penalty

Two common post hoc analytic techniques involve no post hoc error rate penalty. Control for α_{FW} is exercised simply by making subsequent analyses contingent upon significance at a higher level. The difference in the two techniques is in the number of contingencies or filters through which the data must pass.

Fisher least significant difference test. The Fisher least significance difference (LSD) test allows researchers to control for compounding familywise Type I error without paying a penalty. Fisher (1951) initially proposed his test for one-way designs. The Fisher technique involves a significant omnibus *F* which is followed by unrestricted comparisons among means. This technique is basically nothing more than a protected t-test. The approach assumes that only the omnibus *F* test is needed to control for Type I error. It is also possible to apply this logic to factorial designs; no further analysis is performed if the omnibus *F* for the interaction is non-significant. A significant omnibus

F for the interaction, on the other hand, would allow the researcher to perform unrestricted simple comparisons. No comparisons at the level of simple effects are conducted. Thus, with the Fisher approach, there is only one filter or contingency prior to testing pairwise differences; the omnibus F.

The Fisher LSD test has had mixed reviews. Some researchers (Hayter, 1986; Keslman, Games & Rogan, 1980; Ramsey, 1981; Ryan, 1980) believe that it is not a good technique because of its lack of control over compounding familywise error. When unequal variances are paired with unequal n the Fisher test was found to be unreliable (Keppel, 1982; Zwick & Marascuillo, 1984). Others (Hayter, 1986; Keslman, et. al., 1980; Ramsey, 1981; Ryan, 1982) believe the Fisher technique can be useful due to its balance between controlling Type I error and power.

The Keppel no-penalty technique. Keppel's (1991) post hoc approach is similar to Fisher's test in that the researcher pays no error rate penalty. In a two-factor study data must pass through two filters; testing significance of the omnibus F for the interaction and also testing significance of simple effects. Biers and his colleagues refer this to as the filter theory because each analysis is dependent on the previous comparison being significant. That is, if the omnibus *F* is found significant then tests are performed at the level of simple effects. Simple comparisons are only done for those simple effects which are found to be significant. Keppel affirms that this filtering significantly reduces the chance of performing comparisons which are not significant, thereby controlling Type I error without a loss in power.

Controlling Familywise Error at the Simple Effects Level.

These approaches also employ a three step contingency process; omnibus, simple effects, simple comparisons. However, some researchers (e.g. Kirk, 1982) use a technique which controls for familywise error by using a correction (post hoc error rate penalty) that takes place during the analysis of simple effects (Keppel, 1991). This is done by dividing the acceptable familywise error by the number of simple effects to be tested; the resulting number is an adjusted α value which is used as the significance level in the analysis of simple effects. Simple comparisons can be conducted at the .05 significance level after the simple effects are found significant for the adjusted α. However, statisticians differ as to what an acceptable error rate for a family of tests should be and this differentiates the three techniques uncovered in the present study. Generally, there are three techniques which can be used to adjust the overall acceptable error rate; they are described below.

Bonferroni. This approach sets α_{FW} to .05, since the interaction represents one family of statistical tests. Keppel describes simple effects as typically being examined from only one perspective ($A@B_j$ or $B@A_j$). The following is how α (α_B) is computed:

$$\alpha_B = {}^{\alpha_{FW}}\!/_{c} = {}^{.05}\!/_{n_{SE}}$$

where α_B is the adjusted per comparison α.

α_{FW} is the acceptable familywise error rate. With the Bonferroni approach, α_{FW} is set at .05. This is because the interaction is assumed to be the only family.

c is the number of simple effects to be conducted.

Modified Bonferroni. This is a technique which is described by Kirk (1982). He states that the simple effect involves both the main effect and the interaction effect (for example if $B@A_j$ is examined, it contains the main effect of B and the interaction effect of B x A). Furthermore, Kirk assumes that both of these effects must be accounted for by holding $\alpha = .05$ for each of these families when defining the familywise error rate for the simple effect. This means that α_{FW} for the simple effect should be defined as .10 (as a result of adding α_{FW} for the main effect of B to the interaction effect of B x A). Compute the adjusted α (α_{MB}) as follows:

$$\alpha_{MB} = \alpha_{FW} / c = .10 / n_{SE}$$

where α_{MB} is the adjusted per comparison α.

α_{FW} is the acceptable familywise error rate. In the Modified Bonferroni approach, α_{FW} is set at .10 because the interaction and the main effect (e.g. the main effect of B) are assumed to be families.

c is the number of simple effects to be conducted.

Modified Bonferroni-Both. The final technique for identifying the appropriate level for α_{FW} is also described by Kirk (1982). He states that both perspectives of the simple effects ($A@B_j$ and B@Aj) can be studied. The example of a 3 x 3 factorial design would result in a total of six simple effects which could be conducted. In this case, the

post hoc analysis of the interaction would involve three families instead of two (the main effect of A, the main effect of B and the interaction effect of A x B). Like the Modified Bonferroni, the α_{FW} of the three families must be added to result in the adjusted value of $\alpha_{MB\text{-}B}$,

$$\alpha_{MB\text{-}B} = {}^{\alpha_{FW}}\!/_{c} = {}^{.15}\!/_{n_{SE}}$$

where $\alpha_{MB\text{-}B}$ is the adjusted per comparison α.

α_{FW} is the acceptable familywise error rate. With this technique the α_{FW} is set at .15, Because of the two, main effects and the interaction effect.

c is the number of simple effects to be conducted; which will be greater than the other Bonferroni procedures due to the fact that the interaction will be examined from two perspectives.

Even though the correction is assessed at the level of simple effects for each of these approaches, the degree of error rate penalties will differ depending on how a family is defined.

Controlling Familywise Error at the Simple Comparisons Level.

Another way which researchers may control for familywise error is at the level of simple comparisons; pairwise comparison of condition mean. A number of techniques have been created which correct at this level, such as the Tukey test (Winer, 1972) and the Scheffé test (Scheffé, 1953). The current research will look mainly at the Tukey test variations because the Scheffé test has been criticized by some researchers for being

overly conservative (Carmer & Swanson, 1973; Keppel, 1991; Petrinovich & Hardyck, 1969).

Typically, the α_{FW} for the Tukey test is set at .05, and the correction is based on the number of simple comparisons to be conducted. The corrected α_{FW} is used at the level of simple comparisons only if a significant simple effect is found at the .05 level. The following formula is used to determine the critical value for Tukey (F_t):

$$F_t = \frac{(q_t)^2}{2}$$

where q_t is the studentized range statistic's tabled value with the following parameters: The number of means to be compared (k), degrees of freedom for error (k (n-1)), and α_{FW}.

Significance is determined by comparing the obtained F value to the adjusted critical value ,F_t. The adjusted critical F_t is larger than the normal tabled F; this decreases the probability of a Type I error taking place. The Tukey tests which will be used in this study are examined more closely below, however they differ mainly in two ways. First, these approaches have differing degrees of penalty paid at the level of simple comparisons, controlled by k in determining q_t. Second, they differ in whether or not the simple effects are tested for significance.

Tukey-penalty for all possible pairwise comparisons (Tukey Overall). With this Tukey variation, a significant omnibus F for the interaction is followed immediately by the testing of the pairwise comparisons at some adjusted significance level (post hoc error

rate penalty). There is no analysis of simple effects with this approach, however the researcher uses the above formula to compute the adjusted critical value (F_{TOvl}). This results in paying a penalty for all possible pairwise comparisons at the level of simple comparisons.

For a 3 x 3 factorial experiment, the way to compute F_{TOvl} is described below. In this example, a total of nine means would be compared (k = 9). For a sample size of 8 and α_{FW} =.05, the critical value Studentized Range Statistic, q_t (9,63), is approximately 4.55. This means that the adjusted critical value for each simple comparison would be computed in the following manner:

$$F_{TOvl} = \frac{(4.55)^2}{2} = 10.35$$

In this example, the α_{PC} is .00205. The equivalent uncorrected critical value of F is approximately 4.00 (F(1,63) with α .05). It is apparent that the Tukey test is more stringent, and thereby reduces the chance of a Type I error.

Tukey-penalty for a row (Tukey Row). With this variation of the Tukey test, compounding familywise error is controlled by first performing an omnibus F and if significant, then by conducting analyses of simple effects using α = .05. If there are significant results at the level of simple effects, then simple comparisons are made using an adjusted critical value (F_{TRow}). Corrections for the tests at the simple comparison level are made for all pairwise comparisons for a given row (e.g., simple effect) instead of all

possible pairwise comparisons. In this case a family is defined by a row instead of all possible pairwise data comparisons.

Looking again at the 3 x 3 design it is possible to compute the adjusted F_{TRow}. For this version of the Tukey test k = 3 because only 3 means per family would be examined. For a sample size of 8 and $\alpha_{FW} = .05$, the critical value for the Studentized Range Statistic, q_t (3,63), is approximately 3.398. In this case, the adjusted critical value for each simple comparison would be computed in the following way:

$$F_{TRow} = \frac{(3.398)^2}{2} = 5.77$$

The resulting α_{PC} is .01926. This critical value gives a more stringent test of significance than an uncorrected test (where $F_{crit} = 4.00$). The penalty paid using this version of the Tukey test is not as great as that paid by Tukey overall at the level of simple comparisons (be aware that this version uses a filter at the level of simple effects).

Monte Carlo Simulation of Type I and Type II Errors

To fully comprehend Monte Carlo techniques one needs to understand the linear additive model, which illustrates that any score, X, can be produced by the following formula:

$$X = \mu + \alpha_i + \beta_j + \alpha_i\beta_j + E\sigma/ij\ \alpha_i$$

where:

μ is the population mean (constant across all scores),

α_i is the effect of treatment i (the main effect of variable A),

β_j is the effect of treatment j (the main effect of variable B),

$\alpha_i\beta_j$ is the effect of the interaction (factorial combination of A x B), and

$E\sigma$/ij is experimental error, which is random, normally distributed with a mean = 0 and a variance typically set at 1.

Type I error is simulated by entering zeros into all cells in the α_i, β_j, and $\alpha_i\beta_j$ matrices and, randomly generating error with a population with a mean of 0 and a standard deviation of 1. This means that any significant effects would be found due to chance, not any true effect. If the significance level was set at .05, then 5% of the results should be significant due to chance.

Type II error can be simulated by adding treatment effects by entering any non-zero combination for effect for A, B, and A x B. Data, which should show significant results, is subsequently produced consisting of the appropriate effect sizes for the three effects (A, B, A x B). A Type II error has occurred if no significant results are found.

A method which involves varying the probability of making a Type II error was developed by Cohen (1988). This method generates effect sizes of different magnitudes by holding error variance constant while changing the treatment magnitude. The strength of the relationship between the independent and dependent variables is represented by the *f* index (effect size index). The *f* index is computed using the following formulae:

$$f = \frac{\sigma_\mu}{\sigma}$$

and

$$\sigma_\mu = \sqrt{\frac{\sum(\sigma i - \sigma)^2}{k}}$$

where:

μ_i is the mean for a given group in the population

μ is the mean of the population

k is the number of means

σ is the standard deviation of the population

The ratio of the treatment magnitude (σ_μ) to the error variance (σ) results in *f*. It is possible to reduce the formula for the *f* index because data is randomly generated with a mean of 0 and a standard deviation of 1. The reduced formula is:

$$f = \sigma_\mu$$

In addition, because the data is generated so the population mean equals 0, the effect size index formula can be reduced to:

$$f = \sqrt{\frac{\sum(\mu_i)^2}{k}}$$

The above formula makes it easier to identify how the effect size coefficient increases as the difference among means does. In behavioral research the effect sizes are either small, medium, large, or very large with coefficients of .10, .25, .40, and .60 respectively (Cohen, 1988).

Cohen's effect size index deals largely with the standardized range of the population, known as the *d* statistic. The following formula defines *d*:

$$d = \frac{\mu_{max} - \mu_{min}}{\sigma}$$

where

μ_{max} is the largest k mean

μ_{min} is the smallest k mean

When, in randomly generated data, $\sigma = 1$, *d* can be reduced to:

$$d = \mu_{max} - \mu_{min}$$

This specifies the maximum difference among means.

The *d* statistic is a measure of the distribution among treatment means. There are three patterns that a researcher might find, as identified by Cohen:

minimum variability: One mean is located at each extreme, with the others at the midpoint.

intermediate variability: The means are spaced equally over the entire range.

maximum variability: Half of the means fall at each extreme.

The *d* statistic formula depends upon the pattern of variability.

1. *minimum variability*:

$$d = f\sqrt{2k}$$

2. *intermediate variability*:

$$d = 2f\sqrt{\frac{3(k-1)}{k+1}}$$

3. *maximum variability*:

$d = 2f$ (when k is even)

$$d = f\frac{2k}{\sqrt{k^2-1}}$$ (when k is odd)

In this study the above formulae, given the effect size and k, can be used to compute *d*. The *d* statistic indicates the difference between the largest and smallest treatment means. For example, in a 3 x 3 factorial design with a small effect ($f = .10$) and where there is minimum variability among treatment means, the following formula can be used to compute *d*:

$$d = f\sqrt{2k}$$

$$d = .10\sqrt{2(3)} = .24495$$

For any given pattern of variability, *d* must be converted to represent treatment means. For this to be possible, the following restrictions must be followed: 1) For any given row or column, the effects must sum to zero (according to the fixed effects model); 2) the sum of the squared effects divided by k equals *f* squared; 3) across levels, the maximum difference between the smallest and largest means is equal to *d*.

Returning to the 3 x 3 example of the factorial experiment presented above, generation of data with minimum variability among means where the main effect of A is small (*f* = .10), the following matrix is used for randomly generated data:

A1	A2	A3
0.12247	0	-0.12247

The *d* value used to produce the matrix was 0.24495. Examination of the matrix shows that the above restrictions are met.

The matrix for an A x B interaction effect is generated by producing extra coefficients for the first row (simple effect) the same way the main effects are produced. These coefficients are consequently rotated across the levels of A. The result is a 3 x 3 factorial matrix with minimum variability among means and a small interaction effect:

	A1	A2	A3
B1	-0.12247	0	0.12247
B2	0	0.12247	-0.12247
B3	0.12247	-0.12247	0

Once again, the *d* value used to produce this matrix was 0.24495. Examination of this matrix shows that the three restraints are met. A null main effect is assumed for the interaction matrix presented above. When the effect coefficients from the appropriate main effect matrix are added to the effect coefficients in each row of the interaction matrix the result is an interaction matrix where the main effect is not zero.

Previous Biers-Directed Studies

This paper is based on the work from three previous studies which were conducted under the direction of Biers; namely Reising (1993), Brake (1994), and Anthony (1995). Monte Carlo simulation studies were used on factorial ANOVAs of varying sizes. Reising employed a 3 x 3 factorial design, Brake increased to a 3 x 5, and Anthony and Biers used a 5 x 3 arrangement.

Differences Among the Previous Study Designs.

Table 2 illustrates more fully the Reising, Brake, and Anthony and Biers studies. Across the three studies 13 different effect size conditions were employed. Condition 0 was used to test the complete H_0 with a null effect of both A and A x B.

Table 2. Effect Size Combinations for the Main Effect of A and for the Interaction Effect of A x B and Which Conditions Were Ran in Which Study.

Condition	Effect Size of A	Effect Size of A x B	Reising 3 x 3	Brake 3 x 5	Anthony 3 x 3	Anthony 3 x 5	Anthony 5 x 3
0[a]	none	none	yes	yes	yes	yes	yes
1	none	S(.10)	yes	-	yes	yes	yes
2	none	M(.25)	yes	yes	yes	yes	yes
3	none	L(.40)	yes	yes	yes	yes	yes
4	none	VL(.60)	-	yes	yes	yes	yes
5	S	S	yes	-	-	-	-
6	S	M	yes	yes	-	-	-
7	S	L	yes	yes	-	-	-
8	S	VL	-	yes	-	-	-
9	M	S	yes	-	-	-	-
10	M	M	yes	yes	-	-	-
11	M	L	yes	yes	-	-	-
12	M	VL	-	yes	-	-	-

a- Condition 0 was used for αfw with the complete H_0, conditions 1-12 were used for Type II error and for αfw for the partial H_0.

Conditions 1-12 were used to test Type II Error, with the main effect for A added for conditions 5-12. Reising's studies tested everything but very large effect size, and based on those results, Brake chose to use a very large effect size in place of the small. The Anthony and Biers study was different than the previous two. The effect size conditions 5-12 were not utilized in order to avoid possible criticism of the main effect contaminating the results for the interaction effect. The effect size conditions used by Anthony and Biers were designed to concentrate on the pure interaction effect. In addition, Anthony and Biers re-ran the 3 x 5 design to include more tests of the partial H_0. The 3 x 3 design was also re-run because they wanted to replicate the study using a new random number generator.

Tables 3-6 show the differences between the studies in terms of their effect size coefficients. For sake of illustration only large coefficients are used. The tables show how the number of simple comparisons differ across patterns and design variations. The tables indicate the effect size coefficients for each condition, whether or not there was a true (non zero) simple effect, and the number of null and non-null simple comparisons.

Table 3. Reising's Effect Size Matrix, 3 x 3 Design, Large Coefficients.

Pattern 1						
	A1	A2	A3	SE[1]	Null SC[2]	Non-Null[3] SC
B1	0.600	0.000	-0.600	True	0	3
B2	0.000	0.000	0.000	No-True	3	0
B3	-0.600	0.000	0.600	True	0	3
Pattern 3						
	A1	A2	A3	SE[1]	Null SC[2]	Non-Null[3] SC
B1	0.346	0.346	-0.693	True	1	2
B2	0.000	0.000	0.000	No-True	3	0
B3	-0.346	-0.346	0.693	True	1	2

1- Indicates whether the row contains True Simple Effects or No True Simple Effects.
2- Gives the number of Simple Comparisons which take place under the Null H_0.
3- Gives the number of Simple Comparisons which do not take place under the Null H_0

Table 4. Brake's Effect Size Matrix, 3 x 5 Design, Large Coefficients.

Pattern 1						
	A1	A2	A3	SE[1]	Null SC[2]	Non-Null[3] SC
B1	-0.489	0.000	0.489	True	0	3
B2	0.000	0.489	-0.489	True	0	3
B3	0.489	-0.489	0.000	True	0	3
B4	0.000	-0.489	0.489	True	0	3
B5	0.000	0.489	-0.489	True	0	3
Pattern 2						
	A1	A2	A3	SE[1]	Null SC[2]	Non-Null[3] SC
B1	-0.283	-0.283	0.566	True	1	2
B2	-0.283	0.566	-0.283	True	1	2
B3	0.566	-0.283	-0.283	True	1	2
B4	-0.566	0.283	0.283	True	1	2
B5	0.566	-0.283	-0.283	True	1	2

1- Indicates whether the row contains True Simple Effects or No True Simple Effects.
2- Gives the number of Simple Comparisons which take place under the Null H_0.
3- Gives the number of Simple Comparisons which do not take place under the Null H_0

Table 5. Anthony & Biers' Effect Size Matrix, 3 x 5 Design, Large Coefficients.

Pattern 1						
	A1	A2	A3	SE[1]	Null SC[2]	Non-Null[3] SC
B1	0.775	0.000	-0.775	True	0	3
B2	0.000	0.000	0.000	No-True	3	0
B3	0.000	0.000	0.000	No-True	3	0
B4	0.000	0.000	0.000	No-True	3	0
B5	-0.775	0.000	0.775	True	0	3
Pattern 2						
	A1	A2	A3	SE[1]	Null SC[2]	Non-Null[3] SC
B1	-0.447	-0.447	0.894	True	1	2
B2	0.000	0.000	0.000	No-True	3	0
B3	0.000	0.000	0.000	No-True	3	0
B4	0.000	0.000	0.000	No-True	3	0
B5	0.447	0.447	0.894	True	1	2

1- Indicates whether the row contains True Simple Effects or No True Simple Effects.
2- Gives the number of Simple Comparisons which take place under the Null H_0.
3- Gives the number of Simple Comparisons which do not take place under the Null H_0

Table 6. Anthony & Biers' Effect Size Matrix, 5 x 3 Design, Large Coefficients.

Pattern 1								
	A1	A2	A3	A4	A5	SE[1]	Null SC[2]	Non-Null[3] SC
B1	-0.775	0.000	0.000	0.000	0.775	True	3	7
B2	0.000	0.000	0.000	0.000	0.000	No-True	10	0
B3	0.775	0.000	0.000	0.000	-0.775	True	3	7
Pattern 2								
	A1	A2	A3	A4	A5	SE[1]	Null SC[2]	Non-Null[3] SC
B1	-0.693	-0.346	0.000	0.346	0.693	True	0	10
B2	0.000	0.000	0.000	0.000	0.000	No-True	10	0
B3	0.693	0.346	0.000	-0.346	-0.693	True	0	10
Pattern 3								
	A1	A2	A3	A4	A5	SE[1]	Null SC[2]	Non-Null[3] SC
B1	-0.400	-0.400	-0.400	0.600	0.600	True	4	6
B2	0.000	0.000	0.000	0.000	0.000	No-True	10	0
B3	0.400	0.400	0.400	-0.600	-0.600	True	4	6

1- Indicates whether the row contains True Simple Effects or No True Simple Effects.
2- Gives the number of Simple Comparisons which take place under the Null H_0.
3- Gives the number of Simple Comparisons which do not take place under the Null H_0

For example, Table 3 (for pattern 1) has two rows (B1 & B3) with true simple effects (non-null effect sizes) and for those rows there are no null simple pairwise comparisons (tests of partial H_0). There is one row (B2), however, which has three null H_0 comparisons (partial H_0) within the context of no true simple effect. As shown, the rows with true simple effects have fewer null simple comparisons than those with no true simple effects.

By comparing Table 4 to Table 5 it becomes obvious why the conditions were re-run by Anthony and Biers. Brake's original study (Table 4) has very few conditions which tested the partial H_0, and none which tested the partial H_0 in the context of no true effects. By changing the coefficient structure (adding lines with no true simple effects),

Anthony and Biers were able to drastically increase the number of partial H_0 simple comparisons. With the Anthony and Biers 3 x 5 there were a large number of partial H_0 comparisons, and many of these were spread out over the three levels of B within the context of no true effect simple effect. The most effective control method for the Anthony & Biers 3 x 5 study would most likely pay the penalty at the level of simple effects (e.g. Bonferroni) because there are more simple comparisons with no true null simple effect.

On the other hand, the 5 x 3 design has many more partial H_0 simple comparisons which are imbedded within true simple effects. Thus in this case, αfw is probably more effectively controlled by a technique which pays the penalty at the level of simple comparisons (e.g. Tukey Row).

Results of the Previous Biers Directed Studies.

The focus of the studies conducted by Biers and his colleagues (Anthony, 1995; Brake, 1994; Reising, 1993) was on familywise error under the complete null hypothesis. Results indicated that all techniques effectively controlled familywise error under the completely null hypothesis for all four study designs (3 x 3, 3 x 5 Brake, 3 x 5 Anthony and Biers, and 5 x 3).

However, preliminary analyses indicated that familywise error was not controlled under the partial null hypothesis. The only technique that effectively controlled familywise error under the partial null hypothesis was Tukey Overall, but it was too stringent resulting in a loss of power. Other than Tukey Overall, Tukey Row gave the most promising compromise of power and α_{FW} under the partial null hypothesis when

there were a large number of simple comparisons within a simple effect and a fewer number of simple effects (i.e. 5 x 3 Design). Bonferroni gave the best balance when there was more simple effects and fewer simple comparisons within each simple effect (Anthony and Biers 3 x 5 design).

The data generated and the analyses conducted by Biers and his colleagues is incomplete making any conclusions tentative at best. First, the Brake 3 x 5 design was never run under the condition of a small effect size (f = .10) making it only marginally comparable to the data provided by the other designs. Second, the Brake 3 x 5 design data was never analyzed for familywise error under the partial H_0. Furthermore, the 5 x 3 Design (Anthony & Biers) was never analyzed for Type II error.

The Present Study

The purposes of the present study were threefold: (1) to replicate the results of the previous three studies using an improved random generator; (2) to fill in the data and analysis gaps which existed in the three previous studies, and (3) to extend the results to three new techniques. To that end, new data were generated under 4 study designs (3 x 3, 3 x 5 Brake, 3 x 5 Anthony & Biers, 5 x 3) for each combination of 5 interaction effect sizes (0, .10, .25, 40, 60) and 2 sample sizes. The present study focuses on the investigation of familywise error under the partial H_0 since it appears to be a key factor in the choice of control techniques.

The previous results from the preliminary analysis suggest that different control techniques would have to be employed for different designs. However, it would be better from a practitioner's perspective to use one technique in all situations. It is for this

reason that the present study investigates three additional control techniques. First, in the previous research conducted by Biers and his colleagues, only one simple comparison penalty technique has been used-- namely Tukey Row. Perhaps a simple comparison technique with a more severe penalty such as Bonferroni Row might solve the problem with which Tukey Row was associated for the 3 x 5 design. Secondly, a double penalty technique where an error rate penalty is paid at both the simple effects and simple comparison levels could lead to less familywise error under the partial H_0 without too much of a sacrifice in power. Two such techniques are explored here; the Dual Bonferroni and the Dual Modified Bonferroni. These new approaches are described in further detail below.

Bonferroni Row This technique is very similar to the standard Bonferroni approach except for the fact that the number of simple comparisons to be conducted is used for the adjustment instead of the number of simple effects. Therefore:

$$\alpha_{BRow} = \alpha_{FW}/c = .05/c$$

where

α_{BRow} is the adjusted per comparison α

α_{FW} is the acceptable familywise error rate. With this technique α_{FW} is set at .05 because the interaction is assumed to be only one family.

c is the number of simple comparisons to be conducted for the row.

Dual Penalty Techniques. A dual penalty technique is one where the researcher pays a penalty at the level of simple effects, and if significant, pays another penalty at the level of simple comparisons. This means that for the Bonferroni Dual approach the

researcher would first conduct an ordinary Bonferroni control procedure at the level of simple effects. If the results were significant they would be followed by testing the simple comparison using the Bonferroni Row technique. The Modified Bonferroni Dual would work the same way except that both Modified Bonferroni approaches would be used in the place of the Bonferroni procedures. The dual technique approaches are designed to give protection at both levels so that, no matter what the design, Type I Error α_{FW} can be effectively controlled.

CHAPTER II

METHOD

A Monte Carlo simulation computer program was used to generate data for a 3 x 3, 3 x 5, and 5 x 3 between-subjects factorial design in accordance with the procedures outlined in the introduction. The data were created using the linear additive model (see introduction) assuming no main effects of Variables A and B. Error was randomly generated using a random normal number generator with a mean of 0 and a standard deviation of 1. Type I error data under the complete null hypothesis was produced using an interaction effect size of zero (i.e., each cell of the interaction design matrix had a coefficient of zero).

Type I error data under the partial null hypothesis and Type II error were simultaneously generated using interaction effect sizes of .10 (small), .25 (moderate), .40 (large), and .60 (very large). Data for two 3 x 5 designs were produced using two different design matrices —one corresponding to the coefficients for the Brake (1994) study [3 x 5 (B)] and one corresponding to the Anthony (1995) study [3 x 5 (A&B)]. Tables 7-10 show the coefficients for all cells of the design matrix for each of the interaction effect sizes.

Design

For the Type I error database under the complete null hypothesis, data were generated using the null interaction effect size coefficients for each of four study designs (3 x 3, 3 x 5 (B), 3 x 5 (A&B), 5 x 3) using two sample sizes (8, 15). There were 10,000 runs or replications of each condition making a total of 80,000 independent experiments across the four study designs (4 x 2 x 10,000).

For Type II error and Type I error under the partial null hypothesis, four variables were manipulated: (1) study design (3 x 3, 3 x 5(B), 3 x 5(A&B), 5 x 3); (2) effect size of the interaction (.10, .25, .40, .60); (3) pattern of variability of the coefficients within each true simple effect (minimum = Pattern 1, medium = Pattern 2; maximum =Pattern 3); and (4) sample size (8, 15). With a 3 level simple effect (i.e., in a 3 x 3 and 3 x 5 design), the coefficients for patterns 1 and 2 are identical. Thus, for the 3 x 3 and 3 x 5 designs, 10,000 runs were made for each combination of 4 interaction effect sizes, 2 patterns, and 2 sample sizes making a total of 160,000 experiments per design. For the 5 x 3 design, however, there were 10,000 replications for each combination of 4 interaction effect sizes, 3 patterns, and 2 sample sizes or a total of 240,000 experiments.

Table 7
Interaction Effect Size Coefficients for the 3 x 3 Design

Pattern 1 = Pattern 2

ESab = .10

	A1	A2	A3	ESSE
B1	0.15000	0	-0.15000	0.1225
B2	0	0	0	0.0000
B3	-0.15000	0	0.15000	0.1225

ESab = .25

	A1	A2	A3	ESSE
B1	0.37500	0	-0.37500	0.3062
B2	0	0	0	0.0000
B3	-0.37500	0	0.37500	0.3062

ESab = .40

	A1	A2	A3	ESSE
B1	0.60000	0	-0.60000	0.4899
B2	0	0	0	0.0000
B3	-0.60000	0	0.60000	0.4899

ESab = .60

	A1	A2	A3	ESSE
B1	0.90000	0	-0.90000	0.7348
B2	0	0	0	0.0000
B3	0.90000	0	0.90000	0.7348

Pattern 3

ESab = .10

	A1	A2	A3	ESSE
B1	0.08660	0.08660	-0.17320	0.1225
B2	0	0	0	0.0000
B3	-0.08660	-0.08660	0.17320	0.1225

ESab = .25

	A1	A2	A3	ESSE
B1	0.21651	0.21651	-0.43302	0.3062
B2	0	0	0	0.0000
B3	-0.21651	-0.21651	0.43302	0.3062

ESab = .40

	A1	A2	A3	ESSE
B1	0.34646	0.34646	-0.69292	0.4899
B2	0	0	0	0.0000
B3	-0.34646	-0.34646	0.69292	0.4899

ESab = .60

	A1	A2	A3	ESSE
B1	0.56962	0.56962	-1.03924	0.7348
B2	0	0	0	0.0000
B3	-0.56962	-0.56962	1.03924	0.7348

Table 8
Interaction Effect Size Coefficients for the 3 x 5 (B) Design

Pattern 1 = Pattern 2

ESab = .10

	A1	A2	A3	ESSE
B1	-0.12247	0.00000	0.12247	0.1000
B2	0.00000	-0. 12247	-0.12247	0.1000
B3	0.12247	-0. 12247	0.00000	0.1000
B4	0.00000	0. 12247	0.12247	0.1000
B5	0.00000	0. 12247	-0.12247	0.1000

ESab = .25

	A1	A2	A3	ESSE
B1	-0.30619	0.00000	0.30619	0.2500
B2	0.00000	0. 30619	-0.30619	0.2500
B3	0.30619	-0. 30619	0.00000	0.2500
B4	0.00000	-0. 30619	0.30619	0.2500
B5	0.00000	0. 30619	-0.30619	0.2500

ESab = .40

	A1	A2	A3	ESSE
B1	-0.48990	0.00000	0.48990	0.4000
B2	0.00000	0. 48990	-0.48990	0.4000
B3	0.48990	-0. 48990	0.00000	0.4000
B4	0.00000	-0. 48990	0.48990	0.4000
B5	0.00000	0. 48990	-0.48990	0.4000

ESab = .60

	A1	A2	A3	ESSE
B1	-0.73485	0.00000	0. 73485	0.6000
B2	0.00000	0. 73485	-0. 73485	0.6000
B3	0. 73485	-0. 73485	0.00000	0.6000
B4	0.00000	-0. 73485	0. 73485	0.6000
B5	0.00000	0. 73485	-0. 73485	0.6000

Pattern 3

ESab= .10

	A1	A2	A3	ESSE
B1	-0.07071	0.14142	-0.07071	0.1000
B2	0.14142	-0.07071	-0.07071	0.1000
B3	-0.07071	-0.07071	0.14142	0.1000
B4	-0.14142	0.07071	0.07071	0.1000
B5	0.14142	-0.07071	-0.07071	0.1000

ESab = .25

	A1	A2	A3	ESSE
B1	-0.17678	0.35355	-0.17678	0.2500
B2	0.35355	-0.17678	-0.17678	0.2500
B3	-0.17678	-0.17678	0.35355	0.2500
B4	-0.35355	0.17678	0.17678	0.2500
B5	0.35355	-0.17678	-0.17678	0.2500

ESab = .40

	A1	A2	A3	ESSE
B1	-0.28284	-0.28284	0.56568	0.4000
B2	-0.28284	0.56568	-0.28284	0.4000
B3	0.56568	-0.28284	-0.28284	0.4000
B4	-0.56568	0.28284	0.28284	0.4000
B5	0.56568	-0.28284	-0.28284	0.4000

ESab = .60

	A1	A2	A3	ESSE
B1	-0.42426	-0.42426	0.84852	0.6000
B2	-0.42426	0.84852	-0.42426	0.6000
B3	0.84852	-0.42426	-0.42426	0.6000
B4	-0.84852	0.42426	0.42426	0.6000
B5	0.84852	-0.42426	-0.42426	0.6000

Table 9
Interaction Effect Size Coefficients for the 3 x 5 (A&B) Design

Pattern 1 = Pattern 2

ESab = .10

	A1	A2	A3	ESSE
B1	0.19365	0	-0.19365	0.1581
B2	0	0	0	0
B3	0	0	0	0
B4	0	0	0	0
B5	-0.19365	0	0.19365	0.1581

ESab = .25

	A1	A2	A3	ESSE
B1	0.48412	0	-0.48412	0.3953
B2	0	0	0	0
B3	0	0	0	0
B4	0	0	0	0
B5	-0.48412	0	0.48412	0.3953

ESab = .40

	A1	A2	A3	ESSE
B1	0.77460	0.00000	-0.77460	0.6325
B2	0	0	0	0
B3	0	0	0	0
B4	0	0	0	0
B5	-0.77460	0.00000	0.77460	0.6235

ESab = .60

	A1	A2	A3	ESSE
B1	1.16190	0.00000	-1.16190	0.9487
B2	0	0	0	0
B3	0	0	0	0
B4	0	0	0	0
B5	-1.16190	0.00000	1.16190	0.9787

Pattern 3

ESab = .10

	A1	A2	A3	ESSE
B1	-0.11180	-0.11180	0.22360	0.1581
B2	0	0	0	0
B3	0	0	0	0
B4	0	0	0	0
B5	0.11180	0.11180	-0.22360	0.1581

ESab = .25

	A1	A2	A3	ESSE
B1	-0.27951	-0.27951	0.55902	0.3953
B2	0	0	0	0
B3	0	0	0	0
B4	0	0	0	0
B5	0.27951	0.27951	-0.55902	0.3953

ESab = .40

	A1	A2	A3	ESSE
B1	-0.44722	-0.44722	0.89444	0.6325
B2	0	0	0	0
B3	0	0	0	0
B4	0	0	0	0
B5	0.44722	0.44722	-0.89444	0.6325

ESab = .60

	A1	A2	A3	ESSE
B1	-0.67082	-0.67082	1.34164	0.9487
B2	0	0	0	0
B3	0	0	0	0
B4	0	0	0	0
B5	0.67082	0.67082	-1.34164	0.9487

Table 10
Interaction Effect Size Coefficients for the 5 x 3 Design

Pattern 1

Esab = .10

	A1	A2	A3	A4	A5	ESSE
B1	-0.19365	0	0	0	0.19365	0.1225
B2	0	0	0	0	0	0
B3	-0.19365	0	0	0	0.19365	0.1225

ESab = .25

	A1	A2	A3	A4	A5	ESSE
B1	-0.48412	0	0	0	0.48412	0.3062
B2	0	0	0	0	0	0
B3	0.48412	0	0	0	-0.48412	0.3062

ESab = .40

	A1	A2	A3	A4	A5	ESSE
B1	-0.77460	0	0	0	0.77460	0.4899
B2	0	0	0	0	0	0
B3	0.77460	0	0	0	-0.77460	0.4988

ESab = .60

	A1	A2	A3	A4	A5	ESSE
B1	-1.16190	0	0	0	1.16190	0.7348
B2	0	0	0	0	0	0
B3	1.16190	0	0	0	-1.16190	0.7348

Pattern 2

Esab = .10

	A1	A2	A3	A4	A5	ESSE
B1	-0.17320	-0.08660	0	0.08660	0.17320	0.1225
B2	0	0	0	0	0	0
B3	0.17320	0.08660	0	-0.08660	-0.17320	0.1225

ESab = .25

	A1	A2	A3	A4	A5	ESSE
B1	-0.43301	-0.2165	0	0.2165	0.43301	0.3062
B2	0	0	0	0	0	0
B3	0.43301	0.2165	0	-0.2165	-0.43301	0.3062

ESab = .40

	A1	A2	A3	A4	A5	ESSE
B1	-0.69282	-0.34641	0	0.34641	0.69282	0.4899
B2	0	0	0	0	0	0
B3	0.69282	0.34641	0	-0.34641	-0.69282	0.4899

Esab = .60

	A1	A2	A3	A4	A5	ESSE
B1	-1.03922	-0.51961	0	0.51961	1.03922	0.7348
B2	0	0	0	0	0	0
B3	1.03922	0.51961	0	-0.51961	-1.03922	0.7348

Table 10
Interaction Effect Size Coefficients for the 5 x 3 Design (Continued)

Pattern 3

ESab = .10

	A1	A2	A3	A4	A5	ESSE
B1	-0.10000	-0.10000	-0.10000	0.15000	0.15000	0.1225
B2	0	0	0	0	0	0
B3	0.10000	0.10000	0.10000	-0.15000	-0.15000	0.1225

ESab = .25

	A1	A2	A3	A4	A5	ESSE
B1	-0.25000	-0.25000	-0.25000	0.37500	0.37500	0.3062
B2	0	0	0	0	0	0
B3	0.25000	0.25000	0.25000	-0.37500	-0.37500	0.3062

ESab = .40

	A1	A2	A3	A4	A5	ESSE
B1	-0.40000	-0.40000	-0.40000	0.60000	0.60000	0.4889
B2	0	0	0	0	0	0
B3	0.40000	0.40000	0.40000	-0.60000	-0.60000	0.4889

ESab = .60

	A1	A2	A3	A4	A5	ESSE
B1	-0.60000	-0.60000	-0.60000	0.90000	0.90000	0.7348
B2	0	0	0	0	0	0
B3	0.60000	0.60000	0.60000	-0.90000	-0.90000	0.7348

The resulting data were then submitted to a series of programs which conducted the statistical analyses. Ten methods for controlling familywise Type I error were applied to the data in these programs. The control techniques used were: Fisher (FISH), Keppel (KEP), Bonferroni (BON), Modified Bonferroni (MB), Modified Bonferroni Both (MBB), Tukey Row (TROW), Bonferroni Row (BROW), Dual Bonferroni (DBON), Dual Modified Bonferroni (DMB), and Tukey Overall (TOVL). A planned comparisons approach (PLAN) was also applied to the data, where the simple comparisons were tested without any penalties or contingencies. The details of these procedures were described in the introduction.

Procedure

Two Pascal computer programs were written—one to generate and analyze the data and one to calculate Type I error per comparison and familywise and Type II error per comparison. The generation and analysis program first created data using the linear additive model. The polar method for normal deviates (Knuth, 1973) was used to generate error that was normally distributed with a mean of 0 and a standard deviation of 1. The program allowed the user to specify: (1) the design (3 x 3, 3 x 5 (B), 3 x 5 (A&B), 5 x 3), (2) the effect size of the interaction (0.00, 0.10, 0.25, 0.40, 0.60), (2) the pattern of the means for each simple true effect (0, 1, 2, 3), and (4) the sample size (8,15). Based upon the condition specified by the user, the program randomly generated the data for each subject by adding the error for each subject to effect size coefficient for the appropriate cell of the design. (See Tables 7-10 for the effect size coefficients).

The generation and analysis program next performed three statistical analyses on the data generated--an omnibus analysis of variance (ANOVA) for a two way factorial design, an analysis of simple effects from the perspective of A @ B_j , and an analysis of simple comparisons (within each simple effect). The F probabilities for each effect in the omnibus analysis, each simple effect, and each simple comparison were output for later data analysis. For each unique condition of the study, 10,000 experiments were generated and analyzed in 10 runs of 1,000 samples each with each run being stored in a separate data file.

The second Pascal program was used to calculate Type I error, Type I familywise error, and Type 2 error for each of the 11 analysis techniques. Type I and familywise error under the complete null hypothesis were assessed using the conditions where the interaction effect size was zero. In this situation where all differences in effect size coefficients are simultaneously zero, any significant effect represented Type I error - rejection of H_o when in fact it was true. Type II error and Type I error (per comparison and familywise) under the partial null hypothesis were determined from the conditions in which the interaction effect size was greater than zero. Under these conditions, a non-significant difference when the difference in effect size coefficients was non-zero represented a Type II error—failure to reject H_o when it was true. Embedded with the data when the interaction effect size coefficients were non-zero were conditions where the difference in the effect size coefficients was zero. In the conditions where the difference in effect size coefficients was zero, any significant difference represented Type I error under the partial null hypothesis. Familywise error under the complete and partial null hypothesis for each of the 11 techniques was determined by counting up the number

of experiments in which there was at least one Type I error across the simple comparisons.

The second analysis program determined significance by comparing the obtained F probabilities to the criterion (F probability) appropriate for each of the 11 statistical analysis techniques. Table 11 presents the criterion (significance level) used for making a decision about significance at each of the levels of analysis for each of the 11 techniques. Notice how the criterion varies as a function of the design for some of the analysis approaches. Base upon the criterion specified in Table 11, the program counted up the number of Type I, Type I familywise, and Type II errors.

Table 11
Decision Probabilities Associated with Each Control Technique at each Stage of Analysis

APPROACH\ANALYSIS	Omnibus (O)	Simple Effects (SE)			Simple Comparisons (SC)		
Skip ==> Skip ==> SC-NP							
PLAN	-----	-----			0.0500		
O-NP ==> Skip ==> SC-NP							
FISH	0.0500	-----			0.0500		
O-NP ==> SE-NP ==> SC-NP							
KEP	0.0500	0.0500			0.0500		
O-NP ==> SE-P ==> SC-NP		**3 X 3**	**3 X 5**	**5 X 3**			
BON	0.0500	0.0167	0.0100	0.0167	0.0500		
MB	0.0500	0.0333	0.0200	0.0333	0.0500		
MBB	0.0500	0.0250	0.0188	0.0188	0.0500		
O-NP ==> SE-NP ==> SC-P					**3 X 3**	**3 X 5**	**5 X 3**
TROW	0.0500	0.0500			0.0190	0.0190	0.0063
BROW	0.0500	0.0500			0.0167	0.0167	0.0050
O-NP ==> SE-P ==> SC-P		**3 X 3**	**3 X 5**	**5 X 3**	**3 X 3**	**3 X 5**	**5 X 3**
DBON	0.0500	0.0167	0.0100	0.0167	0.0167	0.0167	0.0050
DMB	0.0500	0.0333	0.0200	0.0333	0.0333	0.0333	0.0100
O-NP ==> Skip ==> SC-P		0.0250	0.0188	0.0188	**3 X 3**	**3 X 5**	**5 X 3**
TOVL	0.0500	-----			0.0020	0.0007	0.0007

CHAPTER III

RESULTS

Accuracy of the Generator

The accuracy of the random number generator was determined by analyzing the per comparison Type I probability under the complete null hypothesis for the planned comparison technique (PLAN). The simple effects for the planned comparisons approach were tested directly for significance at the .05 significance level (no omnibus or simple effects tests). The expected Type I error probability, therefore, is .05 for each simple comparison.

There were a total of 69 pairwise simple comparisons in each of the two sample sizes that were tested for significance across the four designs in the study (3 x 3, 3 x 5(B), 3 x 5(A&B) and 5 x 3). Keep in mind that the experiments were generated in ten runs of 1,000 experiments each. Therefore, there were 1,380 experiments (69 x 2 x 100) of 1,000 experiments each for the complete null hypothesis across the four study designs.

Both Table 12 and Figure 4 represent the frequency distribution of the 1,380 cases for Type I error and the planned comparison approach. The probability of Type I error ranged from 0.0300 to 0.0710 with a mean of 0.0501 and a standard deviation of 0.0068. Figure 4 demonstrates that the distribution of Type I error probabilities approximated a normal distribution, as would be expected with a random number generator. These

results indicate that the random number was accurate in estimating Type I error probability.

Table 12 is also helpful for setting boundaries when making a decision about whether the probabilities sufficiently differ from .05 in all ensuing analyses where .05 is the target criterion. Table12 shows that 95% of the runs have a probability value which is less than or equal to the approximate value of .0615 when the true or expected probability is .0500. Therefore, the probability of .0615 will be used when determining when a control technique does not sufficiently control for Type I error per comparison or per familywise.

Table 12.
Plot of frequency distribution of Type I error per comparison under the complete null hypothesis for PLAN

	Frequency	Percent	Cumulative Percent
.0300	1	.1	.1
.0320	3	.2	.3
.0330	3	.2	.5
.0340	1	.1	.6
.0350	8	.6	1.2
.0360	5	.4	1.5
.0370	10	.7	2.2
.0380	18	1.3	3.6
.0390	25	1.8	5.4
.0400	33	2.4	7.8
.0410	35	2.5	10.3
.0420	44	3.2	13.5
.0430	45	3.3	16.7
.0440	60	4.3	21.1
.0450	50	3.6	24.7
.0460	80	5.8	30.5
.0470	75	5.4	35.9
.0480	74	5.4	41.3
.0490	89	6.4	47.8
.0500	82	5.9	53.7
.0510	78	5.7	59.3
.0520	83	6.0	65.4
.0530	65	4.7	70.1
.0540	58	4.2	74.3
.0550	67	4.9	79.1
.0560	51	3.7	82.8
.0570	49	3.6	86.4
.0580	23	1.7	88.0
.0590	44	3.2	91.2
.0600	19	1.4	92.6
.0610	23	1.7	94.3
.0620	20	1.4	95.7
.0630	17	1.2	97.0
.0640	12	.9	97.8
.0650	9	.7	98.5
.0660	7	.5	99.0
.0670	6	.4	99.4
.0680	3	.2	99.6
.0690	3	.2	99.9
.0700	1	.1	99.9
.0710	1	.1	100.0
Total	1380	100.0	

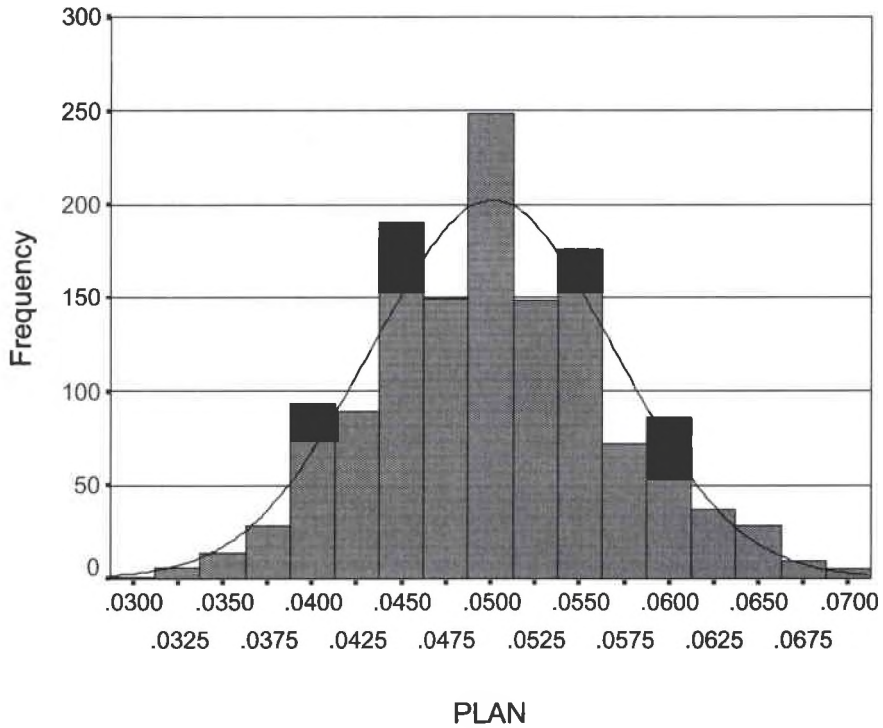

Figure 4. Plot of Frequency Distribution of Type I Error Per Comparison Under the Complete Null Hypothesis for PLAN

Type I Error

Table 13 gives an overview of the control techniques used in the current study. There are seven classes of techniques, which are illustrated in Table 13. The first technique, planned comparisons, is where the simple comparisons are directly tested for significance at the .05 level. Planned comparisons do not attempt to control for α_{FW} and thereby provide a baseline for the comparison of all other techniques. The basic idea of the remaining control techniques is to minimize α_{FW} by doing one of the following: 1) testing the omnibus F (FISH), 2) inserting an additional test of simple effects (KEP), 3) adding a test of simple effects with an error rate penalty which increases with the number of simple effects (BON, MB, MBB), 4) inserting a test for simple comparisons with an

error rate penalty for the number or comparisons within a simple effect (TROW, BROW), 5) utilizing a dual error rate penalty at both the simple effects and simple comparisons level (DBON, DMB), or 6) inserting a test for simple comparisons which pays a penalty for all possible comparisons (TOVL) and which increases as a function of the number of possible comparisons. Table 13 also presents the significance level criterion that is used for making decisions for each level of analysis for each of the 11 techniques. Observe how the criterion varies as a function of the design for some of the analysis techniques.

The Type I error results will be broken down and presented in four sections. Type I error under the complete null hypothesis where all pairwise comparisons are simultaneously zero will addressed in the first two sections. The remaining two sections look at Type I error under the partial null hypothesis, this is where null differences are embedded within context of true differences. With each type of hypothesis both the familywise and per comparison error are discussed.

Table 13.
Decision Probabilities Associated with Each Control Technique at Each Stage of Analysis

APPROACH\ANALYSIS	Omnibus (O)	Simple Effects (SE)			Simple Comparisons (SC)		
Skip ==> Skip ==> SC-NP							
PLAN	-----		-----		0.0500		
O-NP ==> Skip ==> SC-NP							
FISH	0.0500		-----		0.0500		
O-NP ==> SE-NP ==> SC-NP							
KEP	0.0500		0.0500		0.0500		
O-NP ==> SE-P ==> SC-NP		**3 X 3**	**3 X 5**	**5 X 3**			
BON	0.0500	0.0167	0.0100	0.0167	0.0500		
MB	0.0500	0.0333	0.0200	0.0333	0.0500		
MBB	0.0500	0.0250	0.0188	0.0188	0.0500		
O-NP ==> SE-NP ==> SC-P					**3 X 3**	**3 X 5**	**5 X 3**
TROW	0.0500		0.0500		0.0190	0.0190	0.0063
BROW	0.0500		0.0500		0.0167	0.0167	0.0050
O-NP ==> SE-P ==> SC-P		**3 X 3**	**3 X 5**	**5 X 3**	**3 X 3**	**3 X 5**	**5 X 3**
DBON	0.0500	0.0167	0.0100	0.0167	0.0167	0.0167	0.0050
DMB	0.0500	0.0333	0.0200	0.0333	0.0333	0.0333	0.0100
O-NP ==> Skip ==> SC-P		**3 X 3**	**3 X 5**	**5 X 3**	**3 X 3**	**3 X 5**	**5 X 3**
TOVL	0.0500	0.0250	0.0188	0.0188	0.0020	0.0007	0.0007

Complete Null Hypothesis: Type I Error Per Comparison

The expectation is that the Type I error per simple comparison for all techniques will be below .05, considering the probabilities in Table 13. Probability theory states the application of successive tests are expected to result in a combined conditional probability for all techniques which is far below the .05 level expected for the planned approach when the tests are contingent on significance at a prior step and pay an error rate penalty.

The Type I error per comparison under the complete null hypothesis for each of the four study designs are presented in Table 14. This table gives the net (combined) significance level used to assess the significance of each individual simple comparison for the different designs. The per comparison error rate is far below .05 for all techniques except planned comparisons.

Table 14.
Type 1 Error Per Comparison Under the Complete Null Hypothesis

Mean

	DESIGN			
	33	35(A&B)	35(B)	53
PLAN	.0501	.0503	.0499	.0502
FISH	.0121	.0096	.0084	.0083
KEP	.0089	.0068	.0062	.0050
BON	.0049	.0029	.0028	.0028
MB	.0074	.0043	.0040	.0042
MBB	.0062	.0042	.0039	.0031
TROW	.0066	.0051	.0046	.0020
BROW	.0063	.0048	.0045	.0017
DBON	.0039	.0025	.0023	.0012
DMB	.0066	.0040	.0037	.0022
TOVL	.0014	.0006	.0005	.0004

Notice there are two 3 x 5 design replications represented in Table 14; one replication representing the Brake study (3 x 5(B)) and the other representing the Anthony and Biers study (3 x 5(A&B)). The distinction between these two studies is arbitrary because the above data was generated under the complete null hypothesis. The two studies represent independent replications under identical conditions and any difference between the two represents sampling error. Examination of the probabilities for both 3 x 5 designs in Table 14 indicates the greatest difference was .0012, which is slightly greater than 1 in 1,000.

Complete Null Hypothesis: Type I Error Familywise

All pairwise simple comparisons are simultaneously zero (i.e., null) under the complete null hypothesis. When testing individual comparisons, a problem arises; the more comparisons tested for significance the greater the probability that at least one will be found significant due to chance (i.e., α_{FW}). Table 15 indicates that the number of simple comparisons tested for significance changes as the type of design changes. There are 9, 15, and 30 simple comparisons tested for significance in the 3 x 3, 3 x 5, and 5 x 3 designs respectively. If left uncontrolled, therefore, the chance that at least one Type I error will be committed should increase in an orderly fashion from the 3 x 3 to the 3 x 5 to the 5 x 3 design.

Notice that the number of pairwise comparisons in Table 15 are divided into independent (orthogonal) and non-independent (non-orthogonal) comparisons. One could use the formula: $\alpha_{FW} = 1-(1-\alpha_{PC})^c$ to predict the magnitude of α_{FW} if all comparisons were independent. By independent it is meant the significance of one simple comparison is not

related in any way to the significance of another simple comparison. In pairwise comparisons this happens when each comparison involves completely different conditions (e.g., A1 vs. A2 and A3 vs. A4). Mathematically estimating α_{FW} for non-independent comparisons is complex (Keppel, 1991).

Table 15.
Number of Comparisons for the Four Designs under the Complete Null Hypothesis

		3 x 3		3 x 5 (A&B)		3 x 5 (B)		5 x 3		
SE	Indep	3	3	5	5	5	5	3	3	3
SC (Null)	Indep	3	3	5	5	5	5	6	6	6
	Non-Indep	6	6	10	10	10	10	24	24	24
	Total	9	9	15	15	15	15	30	30	30

However different pairwise comparisons involving at least one common condition are non-independent (e.g., A1 vs. A2 and A1 vs. A3). Non-independent comparisons do not increase familywise error as much as independent comparisons. That is, it is more likely that if one redundant comparison is non-significant then others will also be non-significant.

The Type I familywise error rate under the complete null hypothesis is presented in Table 16 for each of the 11 post hoc control techniques. In this table and all that

follow, probabilities that exceed the .0615 criterion will be highlighted in gray. Keep in mind that it was established if the probability exceeded .0615, one could be 95% confident that the probability differed from .05.

Table 16.
Familywise Type 1 Error Rate Under the Complete Null Hypothesis

Mean

	DESIGN			
	33	35(A&B)	35(B)	53
PLAN	.3158	.4734	.4706	.6245
FISH	.0523	.0511	.0457	.0501
KEP	.0436	.0469	.0427	.0390
BON	.0235	.0224	.0212	.0208
MB	.0365	.0331	.0309	.0324
MBB	.0308	.0321	.0298	.0225
TROW	.0414	.0458	.0416	.0342
BROW	.0405	.0443	.0411	.0320
DBON	.0235	.0224	.0212	.0198
DMB	.0365	.0331	.0309	.0316
TOVL	.0108	.0077	.0067	.0102

All techniques, regardless of their control philosophy, except the planned comparison approach, effectively control α_{FW} under the complete null hypothesis. Predictably, the planned comparison approach (which does not control for α_{FW}) has a familywise error rate well above .05 and this increases as a function of the number of comparisons tested for significance (i.e., type of design). Due to the great number of redundant comparisons, the α_{FW} rate of increase for PLAN is not as large as one would expect, if all comparisons were independent. Notwithstanding this fact, there is a 31.58% chance of finding at least one result significant due to chance if α_{FW} is not controlled for, even in the 3 x 3 design.

Partial Null Hypothesis: Type I Error Per Comparison

Most of the data were generated under the condition where there was a true population interaction. The effect size of the interaction varied (small = .10, moderate = .25, large = .40, and very large = .60) under three patterns of significance (P1 = one mean at each extreme with remainder in the middle, P2 = means are equally spaced, P3 = half the means are at each extreme). There were no differences in the coefficients for patterns 1 and 2 for the 3 x 3 and 3 x 5 designs, therefore there were no numbers generated for P2.

Embedded within these true interaction effects were the null simple comparisons. In other words, the true population difference in means was zero. These null simple comparisons cases represent tests of the partial null hypothesis within the context of true interaction effects.

The per comparison Type I error rate is presented in Table 17 for each of the 11 approaches for the effect size of the interaction, type of design, and pattern. First, notice how the probabilities of the PLAN approach are approximately .05, which is where they should be. Second, observe that as the interaction effect size increases (follow the column down for any given technique) Type I error rate for each control technique typically increases. Because the tests for the interaction and simple effects do not provide the amount of protection any longer, this is to be expected since true effects are represented in this condition. The per comparison error rates under the partial null hypothesis were higher than those under the complete null hypothesis (Table 14). The per comparison Type I error rate under the partial null hypothesis is below .05 for all techniques, regardless of the above differences.

Table 17. Type 1 Error Per Comparison under the Partial Null Hypothesis

Mean

		DESIGN							
		33		35(A&B)		35(B)	53		
		P		P		P	P		
ESAB		1	3	1	3	3	1	2	3
Small (.10)	PLAN	.0504	.0492	.0509	.0504	.0502	.0502	.0505	.0493
	FISH	.0170	.0156	.0151	.0152	.0153	.0139	.0136	.0133
	KEP	.0113	.0110	.0099	.0105	.0110	.0082	.0074	.0080
	BON	.0059	.0060	.0033	.0044	.0049	.0045	.0036	.0045
	MB	.0092	.0092	.0055	.0067	.0071	.0067	.0058	.0064
	MBB	.0076	.0079	.0053	.0065	.0070	.0049	.0040	.0048
	TROW	.0082	.0079	.0072	.0073	.0075	.0028	.0027	.0026
	BROW	.0077	.0073	.0067	.0069	.0071	.0024	.0023	.0022
	DBON	.0046	.0046	.0028	.0033	.0037	.0017	.0015	.0017
	DMB	.0083	.0082	.0051	.0059	.0062	.0032	.0029	.0029
	TOVL	.0016	.0014	.0006	.0006	.0006	.0005	.0004	.0006
Moderate (.25)	PLAN	.0494	.0514	.0497	.0496	.0505	.0498	.0485	.0505
	FISH	.0357	.0360	.0384	.0384	.0386	.0373	.0361	.0381
	KEP	.0195	.0251	.0204	.0235	.0314	.0206	.0123	.0225
	BON	.0082	.0151	.0055	.0095	.0169	.0131	.0049	.0146
	MB	.0143	.0208	.0098	.0140	.0225	.0175	.0089	.0194
	MBB	.0113	.0181	.0093	.0135	.0221	.0138	.0055	.0152
	TROW	.0142	.0142	.0144	.0149	.0159	.0048	.0043	.0049
	BROW	.0130	.0128	.0135	.0139	.0144	.0040	.0036	.0041
	DBON	.0065	.0085	.0046	.0059	.0090	.0030	.0020	.0030
	DMB	.0129	.0166	.0090	.0118	.0173	.0060	.0044	.0063
	TOVL	.0018	.0018	.0008	.0007	.0006	.0007	.0005	.0006
Large (.40)	PLAN	.0504	.0497	.0500	.0501	.0509	.0508	.0509	.0499
	FISH	.0476	.0467	.0489	.0490	.0499	.0495	.0494	.0486
	KEP	.0246	.0323	.0245	.0285	.0452	.0281	.0157	.0295
	BON	.0096	.0213	.0059	.0127	.0327	.0206	.0064	.0229
	MB	.0177	.0274	.0109	.0173	.0382	.0247	.0112	.0268
	MBB	.0143	.0247	.0103	.0167	.0377	.0213	.0069	.0234
	TROW	.0173	.0169	.0170	.0170	.0195	.0058	.0054	.0058
	BROW	.0157	.0150	.0158	.0156	.0175	.0047	.0047	.0048
	DBON	.0074	.0098	.0048	.0065	.0142	.0034	.0027	.0036
	DMB	.0160	.0210	.0100	.0138	.0276	.0074	.0058	.0076
	TOVL	.0017	.0019	.0007	.0007	.0007	.0008	.0008	.0007
Very Large (.60)	PLAN	.0500	.0499	.0496	.0503	.0497	.0506	.0510	.0498
	FISH	.0499	.0498	.0496	.0503	.0496	.0506	.0510	.0498
	KEP	.0248	.0352	.0248	.0298	.0491	.0291	.0164	.0307
	BON	.0095	.0260	.0058	.0141	.0449	.0226	.0067	.0254
	MB	.0177	.0311	.0110	.0184	.0472	.0262	.0120	.0282
	MBB	.0135	.0285	.0105	.0179	.0471	.0231	.0074	.0258
	TROW	.0174	.0184	.0169	.0178	.0187	.0058	.0056	.0055
	BROW	.0159	.0164	.0157	.0163	.0166	.0048	.0048	.0045
	DBON	.0071	.0112	.0047	.0071	.0159	.0035	.0027	.0035
	DMB	.0158	.0233	.0101	.0146	.0322	.0074	.0060	.0073
	TOVL	.0021	.0023	.0006	.0007	.0007	.0007	.0007	.0007

Partial Null Hypothesis: Type I Error Familywise

This section will address Type I familywise error under the partial null hypothesis where all simple comparisons are not simultaneously zero. In other words, there is a true interaction effect but some of the differences between the conditions represent null effects.

The effect size of the interaction and the simple effect influence the degree of protection provided. Table 18 represents the relationship between the effect size of the interaction and the effect size of the simple effect for each study design. There will either be a null (0.000) or a true simple effect (some finite number) associated with the effect size of the simple effect (ESSE). A null simple comparison, in other words, can be embedded within context of a null simple effect or a true simple effect. Control techniques at the level of the simple effect should be less likely to work when the null simple comparison is embedded within the context of a true simple effect. Notice the 3 x 5(B) design has no null simple effect.

Table 18 shows that effect size of the simple effect increased as the effect size of the interaction increased. It is worth noting several additional relationships in Table 18 to facilitate later understanding of the results. First, the true simple effect size corresponded exactly to the effect size of the interaction for the 3 x 5(B) study design. Second, in the 3 x 3 and 5 x 3 study designs, the size of the true simple effects were identical and both were slightly larger than the effect size for the interactions. For the 3 x 5(A&B) design, however, the size of the true ESSE was of greater magnitude than the size of the corresponding interaction effects. An ESSE of .3953, for example, was associated with an interaction effect size of .25 (moderate).

Table 18.
Relationship Between the Effect Size of the Interaction and the Effect Size of the Simple Effect.

	Interaction	3 x 3	ESSE 3 x 5(A&B)	3 x 5(B)	5 x 3
Small	0.1000	0.0000 0.1225	0.0000 0.1581	----- 0.1000	0.0000 0.1225
Moderate	0.2500	0.0000 0.3062	0.0000 0.3953	----- 0.2500	0.0000 0.3062
Large	0.4000	0.0000 0.4899	0.0000 0.6325	----- 0.4000	0.0000 0.4899
Very Large	0.6000	0.0000 0.7348	0.0000 0.9487	----- 0.6000	0.0000 0.7348

The results of Type I error familywise under the partial null hypothesis are displayed according to whether or not the null simple comparisons are tested within the context of a null simple effect, within the context of a true simple effect, or overall.

No True Simple Effect.

The study designs where there is no true effect (null effect) are compared in Table 19. The table displays the number of simple effects that are null, and the number of null simple comparisons within the context of these null simple effects that are independent and non-independent. Considering the total number of simple comparisons, the familywise error under the null simple effects is expected to be greater for the 3 x 5 (A&B) and 5 x 3 designs than for the 3 x 3 design. Despite the fact that the total number of simple comparisons is similar for the 3 x 5(A&B) and the 5 x 3 design (9 vs. 10), the

familywise error for the 3 x 5(A&B) design should be greater because there are more independent null simple comparisons (3 vs. 2), (Note: Independent comparisons add more to familywise error than non-independent comparisons do; see previous discussion).

Table 19.
Number of Null Simple Comparisons Embedded Within the Null Simple Effects for Each Study Design when a Null Simple Effect.

		3 x 3		3 x 5 (A&B)		3 x 5 (B)		5 x 3		
		P1	**P3**	**P1**	**P3**	**P1**	**P3**	**P1**	**P2**	**P3**
SE	Indep	1	1	3	3	0	0	1	1	1
SC (Null)	Indep	1	1	3	3	0	0	2	2	2
	Non-Indep	2	2	6	6	0	0	8	8	8
	Total	3	3	9	9	0	0	10	10	10

The results of familywise Type I error under the partial null hypothesis as a function of the effect size of the interaction is presented in Table 20 for each study design. Three points need to be made about Table 20. First, the magnitude of the familywise error behaves as expected (based on Table 19). When the results for PLAN (where the simple comparisons are tested directly) are examined, the familywise error rate was greatest for the 3 x 5 design (approximately .32), slightly less for the 5 x 3 design (.28), and least for the 3 x 3 design (.12). Additionally, the effect size of the interaction had no effect on the probabilities for PLAN. The familywise error under the partial null hypothesis for all other techniques increases with the effect size of the interaction.

Table 20.
Familywise Error Rate Under the Partial Null Hypothesis when a Null Simple Effect.

Mean

		DESIGN						
		33		35(A&B)		53		
		P		P		P		
ESAB		1	3	1	3	1	2	3
Small (.10)	PLAN	.1222	.1203	.3233	.3208	.2812	.2832	.2778
	FISH	.0377	.0365	.0780	.0780	.0626	.0615	.0607
	KEP	.0217	.0209	.0513	.0513	.0228	.0225	.0206
	BON	.0100	.0095	.0164	.0186	.0099	.0097	.0091
	MB	.0165	.0160	.0280	.0304	.0171	.0168	.0147
	MBB	.0134	.0133	.0268	.0294	.0108	.0110	.0098
	TROW	.0201	.0197	.0480	.0483	.0186	.0179	.0165
	BROW	.0193	.0188	.0454	.0469	.0171	.0166	.0149
	DBON	.0100	.0095	.0164	.0186	.0095	.0094	.0087
	DMB	.0165	.0160	.0280	.0304	.0166	.0164	.0142
	TOVL	.0043	.0037	.0055	.0050	.0044	.0039	.0045
Moderate (.25)	PLAN	.1205	.1247	.3176	.3177	.2760	.2815	.2828
	FISH	.0844	.0844	.2370	.2370	.1974	.1999	.2035
	KEP	.0388	.0391	.1125	.1159	.0420	.0395	.0453
	BON	.0139	.0142	.0272	.0262	.0159	.0133	.0154
	MB	.0268	.0266	.0500	.0510	.0298	.0266	.0316
	MBB	.0202	.0200	.0475	.0479	.0175	.0153	.0170
	TROW	.0356	.0350	.1021	.1057	.0324	.0312	.0344
	BROW	.0334	.0327	.0975	.1005	.0285	.0277	.0309
	DBON	.0139	.0142	.0272	.0262	.0150	.0124	.0145
	DMB	.0268	.0266	.0500	.0510	.0283	.0256	.0306
	TOVL	.0047	.0048	.0064	.0061	.0062	.0045	.0060
Large (.40)	PLAN	.1222	.1218	.3199	.3225	.2853	.2879	.2844
	FISH	.1150	.1139	.3120	.3135	.2759	.2776	.2747
	KEP	.0490	.0480	.1391	.1375	.0521	.0497	.0500
	BON	.0168	.0150	.0289	.0272	.0171	.0167	.0174
	MB	.0330	.0315	.0569	.0564	.0341	.0329	.0347
	MBB	.0258	.0231	.0535	.0528	.0197	.0186	.0194
	TROW	.0446	.0427	.1242	.1240	.0398	.0390	.0399
	BROW	.0415	.0399	.1176	.1169	.0360	.0356	.0359
	DBON	.0168	.0150	.0289	.0272	.0162	.0161	.0169
	DMB	.0330	.0315	.0569	.0564	.0330	.0320	.0333
	TOVL	.0051	.0052	.0058	.0055	.0066	.0068	.0067
Very Large (.60)	PLAN	.1230	.1183	.3180	.3227	.2857	.2878	.2837
	FISH	.1227	.1180	.3180	.3226	.2857	.2878	.2836
	KEP	.0500	.0488	.1402	.1425	.0533	.0521	.0496
	BON	.0166	.0162	.0288	.0300	.0176	.0179	.0170
	MB	.0332	.0332	.0571	.0588	.0364	.0354	.0329
	MBB	.0246	.0244	.0539	.0552	.0201	.0201	.0192
	TROW	.0456	.0445	.1245	.1282	.0418	.0410	.0387
	BROW	.0424	.0415	.1178	.1207	.0384	.0369	.0343
	DBON	.0166	.0162	.0288	.0300	.0168	.0171	.0159
	DMB	.0332	.0332	.0571	.0588	.0351	.0341	.0317
	TOVL	.0060	.0063	.0056	.0058	.0062	.0064	.0068

Second, Table 20 indicates that FISH does not sufficiently control familywise error under the partial null hypothesis when there is a null simple effect. Testing the interaction does filter out some of the familywise error. However it is not adequate, especially when the interaction effect size is moderate or greater. This implies that it is necessary to test the simple effects for significance to control familwise error.

Finally, Table 20 shows that with the exception of FISH, KEP, and the simple comparison penalty techniques (TROW and BROW) for the 3 x 5 (A&B) design, all post hoc control techniques adequately control for familywise error under the partial null hypothesis. The simple effects are tested, but no simple effect penalty is paid for KEP, TROW, BROW. These three control techniques control for α_{FW} in all study designs except the 3 x 5(A&B) design. In the 3 x 5(A&B) design the null simple comparisons are not isolated in one simple effect, but are spread across three simple effects (see Table 19). The more simple effects tested for significance, the greater the likelihood of finding at least one significant due to chance. In the 3 x 5(A&B) design it is necessary to pay a simple effect error rate penalty (BON, MB, MBB) to control familywise error at the level of simple effects because the null simple effects are spread across three independent simple effects. For large interaction effect sizes, BON appears to be the most effective of the simple effect error rate penalty techniques.

True Simple Effect

The study designs presented in Table 21 are compared in terms of the partial null hypothesis when there is a true simple effect. When there are true simple effects it is presumed that the simple effect penalty techniques will be less likely to control α_{FW}.

Table 21 indicates that for the 3 x 3, 3 x 5(A&B), and 5 x 3 designs the null simple comparisons are spread across 2 true simple effects, and are spread across 5 true simple effects for the 3 x 5(B) design. The magnitude of familywise error is expected to be larger in the 3 x 5(B) and 5 x 3 designs because the number of null simple comparisons is larger in those designs.

Table 21.
Number of Null Simple Comparisons Embedded Within True Simple Effects for Each Study Design when a True Simple Effect.

		3 x 3		3 x 5 (A&B)		3 x 5 (B)		5 x 3		
		P1	**P3**	**P1**	**P3**	**P1**	**P3**	**P1**	**P2**	**P3**
SE	Indep	0	2	2	2	0	5	2	0	2
SC (Null)	Indep	0	2	0	2	0	5	2	0	4
	Non-Indep	0	0	0	0	0	0	4	0	4
	Total	0	2	0	2	0	5	6	0	8

The results of the Monte Carlo simulation of Type I familywise error under the partial null hypothesis are presented in Table 22. Predictably, the magnitude of familywise error for PLAN is larger for the 3 x 5(B) and 5 x 3 designs and increases as the effect size of the interaction increases. Recall that as the effect size of the interaction increases, the effect size of the simple effect increases (see Table 18).

Table 22 indicates that all post hoc techniques except FISH effectively control familywise error under the partial null hypothesis for small effect sizes. For moderate and larger effect sizes, however, KEP and the simple effect penalty techniques (BON, MB, MBB) do not adequately control Type I familywise error. This is especially true for the 3 x 5(B) and 5 x 3 designs and indicates that simple effects penalty is not adequate for controlling α_{FW} when null simple comparisons are embedded within the context of true interaction effects.

With the exception of the 3 x 5(B) design, the simple comparison penalty techniques (TROW, BROW) and the DBON dual penalty technique all appear to control α_{FW} for all designs. This indicates that it is better to pay a penalty at the level of simple comparisons when there is a null simple comparison embedded within a true simple effect because testing the simple effects for significance does not provide an effective filter any longer. Even the simple comparison penalty is not always sufficient for the 3 x 5 design because the comparisons are spread across five simple comparisons and the amount of penalty paid is not adequate for three comparisons. These techniques do work, on the other hand, for the 3 x 5 design because there are only two simple effects where the null simple comparisons were isolated and a larger penalty was paid for making 10 comparisons within a simple effect.

TOVL is the best technique for controlling Type I familywise error under the partial null hypothesis when there is a true simple effect, as shown by Table 22. The simple effects are not tested in TOVL, rather it involves paying a penalty for testing all pairwise comparisons, regardless of simple effects. Table 6 indicates that TOVL involves comparisons ranging from the .0004 to the .0008 significance level.

Table 22. Familywise Error Rate Under the Partial Null Hypothesis when a True Simple Effect.

Statistics: Mean

ESAB	Variables	DESIGN 33 P 3	35(A&B) P 3	35(B) P 3	53 P 1	53 P 3
Small (.10)	PLAN	.0949	.1010	.2265	.2265	.2996
	FISH	.0280	.0296	.0618	.0548	.0687
	KEP	.0222	.0248	.0489	.0377	.0495
	BON	.0133	.0136	.0240	.0230	.0300
	MB	.0189	.0181	.0338	.0323	.0414
	MBB	.0164	.0177	.0333	.0249	.0316
	TROW	.0146	.0154	.0345	.0149	.0186
	BROW	.0135	.0144	.0330	.0127	.0161
	DBON	.0094	.0085	.0179	.0092	.0127
	DMB	.0162	.0149	.0298	.0171	.0203
	TOVL	.0029	.0010	.0030	.0029	.0042
Moderate (.25)	PLAN	.0988	.0983	.2262	.2296	.3035
	FISH	.0692	.0741	.1699	.1661	.2216
	KEP	.0638	.0710	.1422	.1433	.1919
	BON	.0499	.0556	.0806	.1112	.1498
	MB	.0596	.0634	.1048	.1325	.1780
	MBB	.0556	.0629	.1031	.1155	.1548
	TROW	.0285	.0306	.0761	.0280	.0377
	BROW	.0254	.0279	.0696	.0234	.0309
	DBON	.0226	.0241	.0436	.0214	.0284
	DMB	.0440	.0462	.0826	.0408	.0544
	TOVL	.0040	.0015	.0031	.0036	.0045
Large (.40)	PLAN	.0959	.0953	.2287	.2284	.2979
	FISH	.0901	.0933	.2242	.2220	.2894
	KEP	.0875	.0927	.2051	.2139	.2789
	BON	.0789	.0869	.1526	.1985	.2598
	MB	.0850	.0901	.1760	.2092	.2724
	MBB	.0832	.0899	.1740	.2003	.2617
	TROW	.0338	.0346	.0937	.0336	.0452
	BROW	.0291	.0305	.0844	.0265	.0365
	DBON	.0281	.0298	.0687	.0259	.0360
	DMB	.0587	.0617	.1304	.0528	.0693
	TOVL	.0040	.0016	.0037	.0043	.0055
Very Large (.60)	PLAN	.0995	.0968	.2221	.2255	.3001
	FISH	.0994	.0968	.2221	.2255	.3000
	KEP	.0992	.0968	.2199	.2249	.2995
	BON	.0980	.0966	.2032	.2237	.2978
	MB	.0988	.0967	.2126	.2247	.2989
	MBB	.0985	.0967	.2119	.2238	.2981
	TROW	.0386	.0369	.0894	.0327	.0425
	BROW	.0330	.0329	.0799	.0257	.0334
	DBON	.0330	.0329	.0767	.0257	.0333
	DMB	.0663	.0658	.1494	.0509	.0683
	TOVL	.0046	.0015	.0036	.0036	.0048

Overall

The total number of independent and non-independent null comparisons are shown in Table 23, across the study designs. One would predict, based on Table 23, that the overall magnitude of familywise error under the partial null hypothesis would be the greatest for the 5 x 3 Pattern 1 and 3 designs. Familywise error for the 3 x 5(A&B) Pattern 3 design should also be high.

Table 23. Overall Number of Null Simple Comparisons Under the Partial Null Hypothesis for the Study Designs

		3 x 3		3 x 5 (A&B)		3 x 5 (B)		5 x 3		
		P1	**P3**	**P1**	**P3**	**P1**	**P3**	**P1**	**P2**	**P3**
SE	Indep	1	3	3	5	0	5	3	1	3
SC (Null)	Indep	1	3	3	5	0	5	4	2	6
	Non-Indep	2	2	6	6	0	0	12	8	12
	Total	3	5	9	11	0	5	16	10	18

Table 24 illustrates familywise error under the partial null hypothesis. There is a predictable pattern that α_{FW} follows, as seen in Table 24. As the number of null comparisons increases so does the magnitude of familywise error and works as a function of the effect size of the interaction and the effect size of the simple effect (for all techniques other than PLAN) (see Table 18).

Table 24 clearly shows the only technique that effectively controls α_{FW} under the partial null hypothesis in all circumstances is TOVL. The reason why this technique works is because it employs a very strict significance level (.0004 to.0008). The high familywise error rates of the FISH show the opposite end of the spectrum. This shows that simply testing the significance of the interaction does nothing to control familywise error under the partial null hypothesis.

With the exception of FISH, all post hoc techniques were able to effectively control familywise error under the partial null hypothesis for the 3 x 3 Pattern 1 and the 5 x 3 Pattern 2 designs. The null simple comparisons in both of these designs are concentrated within one simple effect (see Table 23). When there is only one simple effect with null simple comparisons familywise can be effectively controlled by testing the simple effect for significance with a penalty (BON, MB, MBB, DBON, DMB). Testing the simple effect with no penalty also worked in this situation if it is followed by simple comparisons with a penalty (TROW, BROW). Familywise error under the partial null is even controlled by KEP, which pays no penalty, except when the interaction effect size is large or very large.

The simple effect penalty techniques (BON, MB, MBB) controlled familywise error for the Pattern 1, 3 x 5(A&B) design, even with moderate to very large interaction effect sizes. The simple comparison penalty techniques (TROW, BROW) did not, on the other hand, control α_{FW} in this condition. Here, the null simple comparisons were spread across three simple effects and not isolated within a single simple effect. Paying a penalty at the level of simple effects in this situation is essential, and paying a penalty at

Table 24. Overall Familywise Error Rate Under the Partial Null Hypothesis

Mean

		DESIGN							
		33		35(A&B)		35(B)	53		
		P		P		P	P		
ESAB		1	3	1	3	3	1	2	3
Small (.10)	PLAN	.1222	.2018	.3233	.3883	.2265	.4409	.2832	.4891
	FISH	.0377	.0570	.0780	.0905	.0618	.0912	.0615	.0941
	KEP	.0217	.0408	.0513	.0686	.0489	.0557	.0225	.0644
	BON	.0100	.0223	.0164	.0317	.0240	.0319	.0097	.0378
	MB	.0165	.0337	.0280	.0465	.0338	.0468	.0168	.0529
	MBB	.0134	.0289	.0268	.0452	.0333	.0347	.0110	.0399
	TROW	.0201	.0332	.0480	.0593	.0345	.0321	.0179	.0335
	BROW	.0193	.0313	.0454	.0574	.0330	.0287	.0166	.0300
	DBON	.0100	.0185	.0164	.0268	.0179	.0184	.0094	.0212
	DMB	.0165	.0312	.0280	.0436	.0298	.0324	.0164	.0331
	TOVL	.0043	.0065	.0055	.0060	.0030	.0072	.0039	.0087
Moderate (.25)	PLAN	.1205	.2094	.3176	.3842	.2262	.4382	.2815	.4946
	FISH	.0844	.1420	.2370	.2834	.1699	.3086	.1999	.3496
	KEP	.0388	.0981	.1125	.1740	.1422	.1759	.0395	.2245
	BON	.0139	.0629	.0272	.0800	.0806	.1245	.0133	.1624
	MB	.0268	.0835	.0500	.1098	.1048	.1566	.0266	.2019
	MBB	.0202	.0737	.0475	.1066	.1031	.1302	.0153	.1683
	TROW	.0356	.0615	.1021	.1314	.0761	.0589	.0312	.0698
	BROW	.0334	.0563	.0975	.1243	.0696	.0509	.0277	.0605
	DBON	.0139	.0359	.0272	.0495	.0436	.0360	.0124	.0425
	DMB	.0268	.0685	.0500	.0937	.0826	.0675	.0256	.0823
	TOVL	.0047	.0087	.0064	.0077	.0031	.0098	.0045	.0105
Large (.40)	PLAN	.1222	.2044	.3199	.3864	.2287	.4451	.2879	.4942
	FISH	.1150	.1910	.3120	.3756	.2242	.4305	.2776	.4770
	KEP	.0490	.1297	.1391	.2166	.2051	.2528	.0497	.3135
	BON	.0168	.0924	.0289	.1113	.1526	.2113	.0167	.2720
	MB	.0330	.1130	.0569	.1412	.1760	.2345	.0329	.2966
	MBB	.0258	.1038	.0535	.1378	.1740	.2150	.0186	.2753
	TROW	.0446	.0744	.1242	.1537	.0937	.0719	.0390	.0828
	BROW	.0415	.0674	.1176	.1431	.0844	.0614	.0356	.0706
	DBON	.0168	.0424	.0289	.0558	.0687	.0420	.0161	.0519
	DMB	.0330	.0876	.0569	.1146	.1304	.0838	.0320	.0991
	TOVL	.0051	.0091	.0058	.0071	.0037	.0109	.0068	.0121
Very Large (.60)	PLAN	.1230	.2045	.3180	.3859	.2221	.4424	.2878	.4941
	FISH	.1227	.2042	.3180	.3858	.2221	.4424	.2878	.4940
	KEP	.0500	.1427	.1402	.2236	.2199	.2646	.0973	.3317
	BON	.0166	.1123	.0288	.1226	.2032	.2369	.0179	.3083
	MB	.0332	.1285	.0571	.1492	.2126	.2518	.0354	.3202
	MBB	.0246	.1202	.0539	.1457	.2119	.2389	.0201	.3098
	TROW	.0456	.0811	.1245	.1597	.0894	.0731	.0410	.0790
	BROW	.0424	.0730	.1178	.1489	.0799	.0630	.0369	.0660
	DBON	.0166	.0485	.0288	.0612	.0767	.0420	.0171	.0485
	DMB	.0332	.0970	.0571	.1202	.1494	.0838	.0341	.0973
	TOVL	.0060	.0109	.0056	.0073	.0036	.0098	.0064	.0115

the level of simple comparisons is not sufficient to control familywise error in this situation.

For the 3 x 5(A&B) and the 3 x 5(B) Pattern 3 designs for moderate to very large interaction effect sizes TOVL was the only technique which adequately controlled familywise error. For the 3 x 5 designs, the null simple comparisons are spread across 5 simple effects, some of which were true simple effects. Even the simple effect penalty techniques were insufficient to control familywise error under the partial null hypothesis in this situation.

DBON is the only other technique other than TOVL that shows promise in controlling α_{FW}. Even DBON falls apart, however, with large and very large interaction effect sizes when the null simple comparisons are embedded in true simple effects and are spread across 5 simple effects (3 x 5(A&B) Pattern 3 and 3 x 5(B) Pattern 3 designs.

To summarize, when the interaction effect sizes are moderate or larger, it appears that the most effective techniques to control familywise error under the partial null hypothesis depend on the number of simple effects that contain null simple comparisons. If there is one simple effect where the null simple effects are isolated, then almost any technique will effectively control familywise error. However, when the null simple comparisons are spread over three simple effects, then using a simple effect penalty technique seems to be the best choice, with BON being the most effective. Finally, when the null simple comparisons are spread over 5 simple effects TOVL, a very conservative test, is the only technique that is consistently effective at controlling familywise error.

Type II Error

Type II error occurs when there is a true treatment effect (i.e., H_0 true), but the results are attributed to sampling error (i.e., fail to reject H_0). Type II error was generated by using a combination of four interaction effect sizes, two effect size patterns (3 for the 5 x 3 design), and two sample sizes for each of the four study designs.

The three statistical elements that have an effect on Type II error for each simple comparison are: (1) effect size of the simple comparison (ESSC), (2) sample size, and (3) error variance. The population error variance was held at a constant 1.0 for the current study. The only things that varied were the sample size ($n = 8$ or $n = 15$) and the effect size of the simple comparison. By directly manipulating the effect size of the simple effect (ESSE), the effect size of the interaction, the pattern, and the design type, it was possible to indirectly vary the ESSC (see the introduction for examples of effect size calculations). Note the effect size of the interaction and the effect size of the simple effect was positively correlated with the ESSC. The intricacy of the Type II error data presentation can be reduced by only considering the ESSC because that effect combines the effects of the study design, the effect size of the interaction, the ESSE, and pattern.

Only selected simple comparison effect sizes were closely examined in order to reduce the complexity of the Type II error rates to a manageable level. Four effect sizes were chosen corresponding to small (.10), moderate (.25), large (.40), and very large (.60) effect sizes. Table 25 presents cases in the Type II error base where the effect sizes (ESSC) were within +/- .05 of the aforementioned values.

Table 25. Type II Error for 11 Control Techniques as a Function of Selected Simple Comparison Effect Sizes and Sample Sizes

	N							
	8				15			
	ESSC				ESSC			
	.10	.25	.40	.60	.10	.25	.40	.60
PLAN	.9298	.8314	.6460	.3354	.9111	.7206	.4130	.0947
FISH	.9780	.8661	.6767	.3410	.9632	.7376	.4279	.0955
KEP	.9839	.8848	.7102	.3709	.9723	.7622	.4566	.1053
BON	.9905	.9170	.7789	.4508	.9828	.8113	.5260	.1411
MB	.9870	.9010	.7444	.4101	.9773	.7867	.4905	.1203
MBB	.9890	.9069	.7553	.4185	.9804	.7957	.4997	.1239
TROW	.9915	.9429	.8180	.5570	.9861	.8771	.6189	.2382
BROW	.9922	.9479	.8301	.5786	.9872	.8859	.6350	.2547
DBON	.9950	.9608	.8618	.6263	.9915	.9051	.6731	.2787
DMB	.9916	.9369	.8004	.5296	.9859	.8577	.5804	.2027
TOVL	.9986	.9911	.9609	.8418	.9976	.9769	.8810	.5514

Table 25 displays the Type II error for each of the 11 control techniques for each sample size and the selected effect sizes. Predictably, the Type II error rates decrease as a function of effect size, this decrease being higher for the larger sample size. The technique with the lowest Type II error rate is the planned comparison approach because it directly tests each simple effect without any contingency. Tukey Overall (TOVL), on the other hand, has the highest Type II error rate because it pays a very strict penalty at the level of simple comparisons.

More importantly, Table 25 demonstrates that the differences in Type II error within the 10 post hoc control techniques increases as a function of the chosen ESSC's (within the selected range) for both sample sizes. The probability of making a Type II error is great (approximately .97 - .99) for small effect sizes (.10), and the difference between the post hoc control techniques is very low (approximately .02 - .03). With very large effect sizes (.60), however, the differences in the size of Type II error within the 10

post hoc control techniques is approximately .50 for n = 8 and .45 for n = 15. In other words, a researcher has a 50% greater chance of making a Type II error using TOVL than if he/she were to use FISH for a sample size of 8.

The complexity of presenting the data for Type II errors can be further reduced if it can be shown that there are minimal differences between similar control techniques. It appeared reasonable to use the .05 criteria to determine the significance in general. Because the data is subject to sampling variability, the same criteria as applied in Type I error assessment was used – if the techniques differed by more than .0615 in Type II error, one could be 95% confident that the control techniques differed by more than 5%. Additional comparisons in Table 25 should be considered in this regard:

1. The techniques that pay only a simple effects rate penalty (BON, MB, MBB) differ slightly, with the greatest difference being .04 between BON and MB when the ESSC = .6 and n = 8. For the designs selected for this study, therefore, there appears to be little difference in the likelihood of experiencing a Type II error among these procedures. Because the data for the partial familywise Type I error indicates the Bonferroni technique is the most promising of the three, BON was selected for further comparison with the other categories of control procedures.
2. The techniques that pay a penalty at the level of simple comparisons (TROW and BROW) differ minimally in Type II error, with 2 % being the largest difference, when the ESSC = .60 and n = 8. Because this difference in Type II error was so slight, the BROW procedure was chosen for other techniques since it allowed a direct comparison with BON.

3 When the dual penalty techniques were compared the results showed that the DBON experiences 6.15 – 9.67% more Type II error than DMB for large and very large effect sizes. This indicates that there is a potentially serious difference in Type II error. This does indicate that both techniques should be further investigated, however, only DBON will be used in subsequent comparisons because the familywise error data under the null hypothesis suggests that it is the more practicable approach.

Table 26 shows the pairwise differences in Type II error for the seven control procedures, with one technique representing each class. Recall that the seven classes are: (1) a direct simple comparison test – PLAN, (2) an omnibus F contingency – FISH, (3) an omnibus F and a simple effects contingency with no simple effects error rate penalty – KEP, (4) an omnibus F and a simple effects contingency, and a simple comparisons error rate penalty – BON, (5) an omnibus F and a simple effects contingency with a dual error rate penalty for simple effects and simple comparisons - BROW, (6) an omnibus F and a simple effects contingency with a dual error rate penalty at both the level of simple effects and simple comparisons – DBON, and (7) an omnibus F contingency and a simple comparisons penalty for all possible pairwise comparisons – TOVL.

Table 26 is arranged by sample size and effect size of the simple comparison only for moderate, large, and very large effect sizes. Small effect sizes were not presented because the differences for the post hoc control techniques were miniscule (approximately .02 - .03). Any differences in Table 26 that exceeded the .615 criterion are highlighted in gray.

Table 26. Differences in Type II Error as a Function of Effect Size of the Simple and Sample Size for Selected Control Techniques

		ESSC = .25					
		FISH	KEP	BON	BROW	DBON	TOVL
n = 8		0.8661	0.8848	0.9170	0.9479	0.9608	0.9911
PLAN	0.8314	-0.0346	-0.0534	-0.0856	-0.1165	-0.1293	-0.1597
FISH	0.8661		-0.0187	-0.0509	-0.0819	-0.0947	-0.1250
KEP	0.8848			-0.0322	-0.0631	-0.0760	-0.1063
BON	0.9170				-0.0310	-0.0438	-0.0741
BROW	0.9479					-0.0128	-0.0432
DBON	0.9608						-0.0303
		FISH	KEP	BON	BROW	DBON	TOVL
n = 15		0.7376	0.7622	0.8113	0.8859	0.9051	0.9769
PLAN	0.7206	-0.0170	-0.0416	-0.0907	-0.1653	-0.1845	-0.2563
FISH	0.7376		-0.0246	-0.0737	-0.1483	-0.1675	-0.2393
KEP	0.7622			-0.0491	-0.1237	-0.1429	-0.2147
BON	0.8113				-0.0746	-0.0938	-0.1656
BROW	0.8859					-0.0192	-0.0910
DBON	0.9051						-0.0718
		ESSC = .40					
		FISH	KEP	BON	BROW	DBON	TOVL
n = 8		0.6767	0.7102	0.7789	0.8301	0.8618	0.9609
PLAN	0.6460	-0.0307	-0.0642	-0.1329	-0.1841	-0.2158	-0.3149
FISH	0.6767		-0.0335	-0.1022	-0.1534	-0.1851	-0.2842
KEP	0.7102			-0.0687	-0.1199	-0.1516	-0.2507
BON	0.7789				-0.0512	-0.0829	-0.1820
BROW	0.8301					-0.0317	-0.1308
DBON	0.8618						-0.0991
		FISH	KEP	BON	BROW	DBON	TOVL
n = 15		0.4279	0.4566	0.5260	0.6350	0.6731	0.8810
PLAN	0.4130	-0.0149	-0.0436	-0.1130	-0.2220	-0.2601	-0.4680
FISH	0.4279		-0.0287	-0.0981	-0.2071	-0.2452	-0.4531
KEP	0.4566			-0.0694	-0.1784	-0.2165	-0.4244
BON	0.5260				-0.1090	-0.1471	-0.3550
BROW	0.6350					-0.0381	-0.2460
DBON	0.6731						-0.2079
		ESSC = .60					
		FISH	KEP	BON	BROW	DBON	TOVL
n = 8		0.3410	0.3709	0.4508	0.5786	0.6263	0.8418
PLAN	0.3354	-0.0056	-0.0355	-0.1154	-0.2432	-0.2909	-0.5064
FISH	0.3410		-0.0299	-0.1098	-0.2376	-0.2853	-0.5008
KEP	0.3709			-0.0799	-0.2077	-0.2554	-0.4709
BON	0.4508				-0.1278	-0.1755	-0.3910
BROW	0.5786					-0.0477	-0.2632
DBON	0.6263						-0.2155
		FISH	KEP	BON	BROW	DBON	TOVL
n = 15		0.0955	0.1053	0.1411	0.2547	0.2787	0.5514
PLAN	0.0947	-0.0008	-0.0106	-0.0464	-0.1600	-0.1840	-0.4567
FISH	0.0955		-0.0098	-0.0456	-0.1592	-0.1832	-0.4559
KEP	0.1053			-0.0358	-0.1494	-0.1734	-0.4461
BON	0.1411				-0.1136	-0.1376	-0.4103
BROW	0.2547					-0.0240	-0.2967
DBON	0.2787						-0.2727

When analyzing Table 26, the following are important comparisons to consider:

1. Fisher (FISH) vs. Keppel(KEP). These techniques demonstrate how Type II error is affected when a simple effects test is used as a contingency before any simple comparisons are performed. Table 26 indicates the added simple effects test (KEP) does not substantially affect Type II error because the largest difference between the two techniques is .033, which is well below the .615 limit.
2. Keppel (KEP) vs. Bonferroni (BON). When the ESSC is large (.40) with both sample sizes and for a small sample size (n = 8) with a very large effect size (.60) the Type II error rates for these two procedures differs by more than 5%. When this occurs, BON is 6.87 – 7.99% more likely to result in a Type II error. This suggests that under certain conditions the Bonferroni error rate penalty has an adverse effect on Type II error at the level of simple effects.
3. Keppel (KEP) vs. Bonferroni Row (BROW). This comparison shows how a penalty at the level of simple comparisons can affect Type II error. The BROW Type II error rate exceeds KEP by more than 5% in all cases, with the difference varying from .063 to .208. This demonstrates that a penalty at the level of simple comparisons affects Type II error more than paying no penalty does.
4. Bonferroni (BON) vs. Bonferroni Row (BROW). These are both penalty techniques, but differ because Bonferroni pays a penalty at the level of

simple effects and Bonferroni Row pays a penalty at the level of simple comparisons. When the sample size is large (n = 15), Bonferroni experiences 7.46 – 11.36 % more type II error than BROW, which exceeds the .615 criteria for all effect sizes. When the effect size is smaller the differences in Type II error (.1278) surpasses the criterion only with a very large effect size (.60). Therefore, there is evidence that a penalty at the level of simple effects can yield fewer Type II errors than a penalty at the level of simple comparisons.

5. Bonferroni (BON) vs. Dual Bonferroni (DBON). This is a comparison of a dual penalty technique and a simple effect penalty technique. The two techniques differed substantially in every instance, except when there was a small sample size and a moderate ESSC. In every other situation, Dual Bonferroni had a greater Type II error rate, which ranged from .0829 (n = 8, ESSC = .40) to .1755 (n = 8, ESSC = .60).

6. Bonferroni Row (BROW) vs. Dual Bonferroni (DBON). The greatest difference between the dual penalty technique (DBON) and the simple comparison penalty technique (BROW) is .048 for a small sample size (n = 8) and a very large effect size (.60). There is no difference between these two techniques that exceed 5%.

7. Tukey Overall (TOVL) vs. All Other Techniques. It is obvious that Tukey Overall has a very high Type II error rate, especially when it is compared to the other control procedures. The only times when TOVL is not substantially different is when it is compared to BROW and DBON for a

small sample size (n = 8) with a moderate effect size (.40). In those situations the differences in Type II error are .0432 and .0303, respectively. In every other case TOVL has a Type II error rate that ranges from .0718 (DBON, n = 15, ESSC = .25) to .5008 (FISH, n = 8, ESSC = .60) higher than other procedures.

In conclusion, merely testing the simple effects for significance will not substantially affect the magnitude of Type II error. If the simple effects are tested as a contingency, Type II error can be affected in some cases (KEP vs. BON) if paying a penalty at the level of simple effects. Paying a penalty at the level of simple comparisons, however, will have a greater impact on Type II error than does paying a penalty at the level of simple effects (KEP vs. BON, KEP vs. BROW, and BON vs. BROW). A dual error rate penalty will result in higher Type II error than a simple effects penalty (BON), but not more than a simple comparisons penalty technique (BROW). It is obvious that Tukey Overall results in the highest Type II error.

Power

Power curves were created for each of the seven control procedures to present the Type II error results in a more usable format. Power is the probability of finding a true treatment effect significant (i.e., rejecting H_0 when it is actually false). Power is equal to $1 - \beta$ where β is the probability of a Type II error.

Smoothed power curves were created for each of the chosen control procedures by curve fitting (i.e., regressing) the transformed Type II error data ($1 - \beta$) to the ESSC for

each sample size. A number of different curve fitting functions were used. In every case, the best fitting line was obtained using a regression equation with a cubic fit. The R-squares ranged from .918 (BON, n = 8) to .997 (PLAN, n = 8) with the median for the control techniques being .974.

Figures 5 and 6 display the smoothed power curves for each of the seven control techniques as a function of the effect size of the simple comparisons for small and large sample sizes. The power curves serve to visually verify the results of the Type II error analysis. The techniques fall into groups. The differences in power between FISH and Keppel are slight, as are those between BROW and DBON. The large jumps in power variances appear to be between KEP and BON, between BON and BROW/DBON, and between BROW/DBON and TOVL. Therefore, the simple effect penalty procedures (BON) are not as powerful as a simple effect test with no penalty (KEP). Techniques that pay a penalty at the level of simple comparisons (BROW) result in a higher power loss than techniques that pay a penalty at the level of simple effects (BON). A dual error rate penalty (DBON) does not lead to any loss of power in comparison to a technique that pays a penalty at the level of simple comparisons (BROW). Finally, Tukey Overall, which pays a penalty for all pairwise comparisons greatly reduces the power to identify true treatment effects, even with effects that are very large or larger.

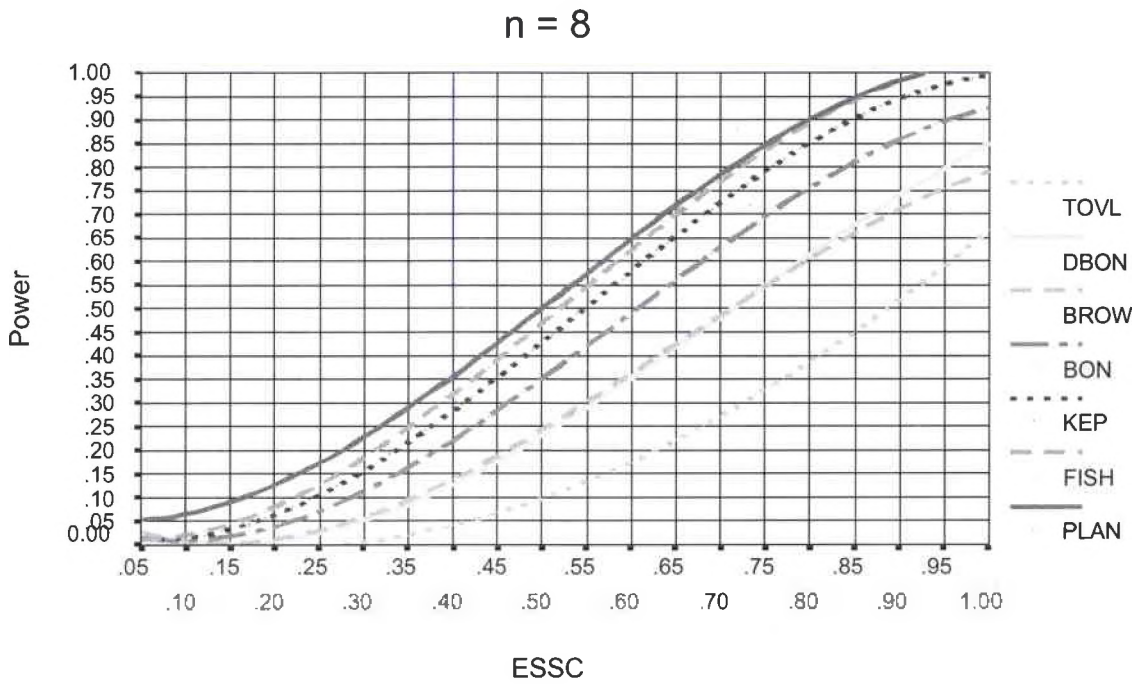

Figure 5. Smoothed Power Curve for Each of the Seven Control Techniques as a Function of the Effect Size of the Simple Comparisons for Small Sample Size

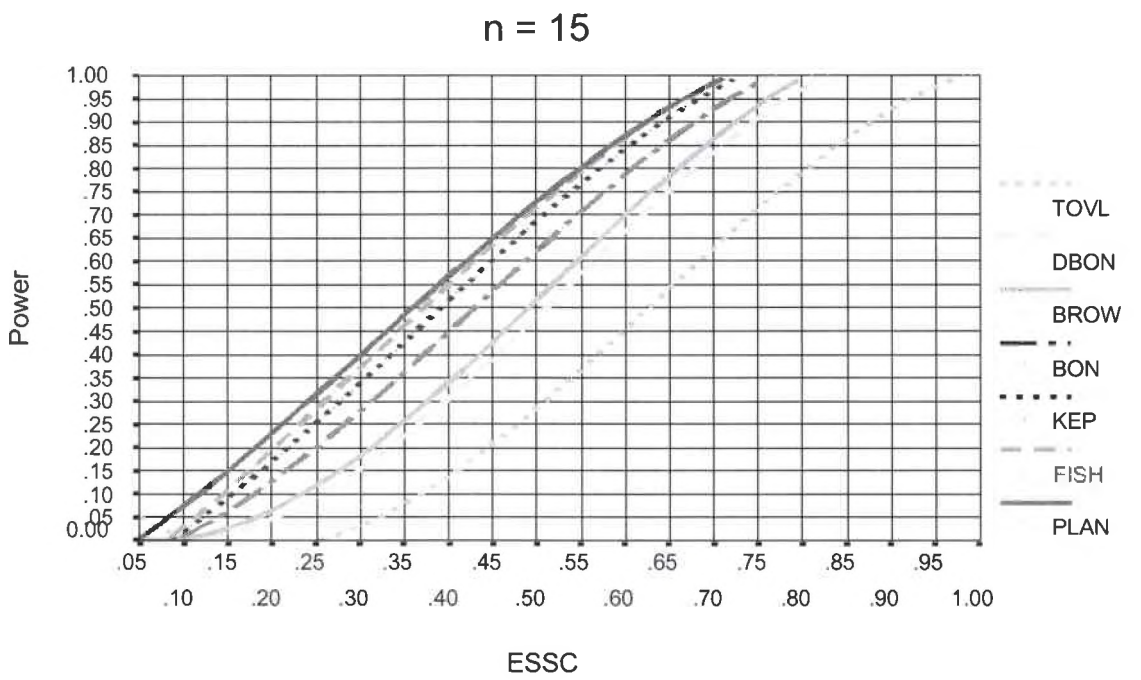

Figure 6. Smoothed Power Curve for Each of the Seven Control Techniques as a Function of the Effect Size of the Simple Comparisons for Large Sample Size

CHAPTER IV

DISCUSSION

The main purpose of the current study was to examine 10 post hoc techniques for controlling familywise error rate in tests of simple comparisons for factorial designs. Familywise error and Type II error under both the complete and partial null hypothesis were investigated for a 3 x 3, 3 x 5, and 5 x 3 factorial design. The techniques minimized familywise error by doing one of the following: 1) testing the omnibus F (Fisher), 2) inserting an additional test of simple effects (Keppel), 3) using a test of simple effects with an error rate penalty for the number of simple effects (Bonferroni, Modified Bonferroni, Modified Bonferroni Both), 4) utilizing a test for simple comparisons with an error rate penalty for the number of comparisons within a simple effect (Tukey Row, Bonferroni Row), 5) inserting a dual error rate penalty at both the simple effects and simple comparisons level (Dual Bonferroni, Dual Modified Bonferroni, or 6) utilizing a test for simple comparisons which pays a penalty for all possible comparisons and which increases as a function of the number of possible pairwise comparisons (Tukey Overall).

Two points of caution must be made regarding the interpretation of familywise error rates in the current study before the results of the control techniques are discussed. The first point is that all pairwise comparisons conducted in the present study were tested only within a simple effect and from one perspective (i.e., A @ bj). The use of this

analysis method reduced familywise error rates because it limited the total number of pairwise comparisons tested for significance in any design. In the 3 x 5 design, for example, this reduced the number of pairwise comparisons from 105 to 30, greatly reducing familywise error. Therefore, this constraint must be kept in mind when interpreting the familywise error results. Had the data not been limited in this way, even fewer post hoc techniques would have been acceptable.

The second point is that the .05 probability level was used as the criterion for the acceptability of any technique when assessing familywise error. That is, the decision could be made that there was no problem with a technique if one could be 95% confident that the familywise error was not greater than .05 (using the .0615 probability). See page 46 for an explanation of the .0615 probability. This indicated that a researcher is willing to tolerate 5% familywise error before a technique is deemed unacceptable. There is no universal criteria for familywise error, unlike the criteria for testing effects for significance. Most researchers use .05 as the criterion, but this is not universal. There are researchers who have adopted the .10 or .15 level, or even multiple criteria depending on the number of conditions examined (see Stevens, 1986). The conclusions about the acceptability of the control techniques in this study are based on the .05 criterion. If one wishes to adopt a different criterion (i.e., .10 or .15), then the conclusions reached in this thesis would be different.

In summary, results of the current study replicated the results of Reising (1993), Brake (1994), and Anthony (1995) using a new random number generator and extended the conclusions to three new techniques. The results of the ten post hoc analysis

techniques are discussed separately below for the complete (classic statistical theory) and partial null hypothesis.

Classic Hypothesis Testing

Classic hypothesis testing under the complete H_0 assumes that all pairwise differences among means are simultaneously zero. Results showed that all control techniques effectively controlled α_{FW} under the complete null hypothesis, regardless of control philosophy or type of design. Fisher was the only questionable technique because the familywise probabilities were all around .05.

Post hoc analysis procedures should be chosen according to Type II error and power rates if the complete null hypothesis is assumed to be true. If Fisher is eliminated (because of marginal control of α_{FW} and difficulties under conditions where ANOVA assumptions are not met –see Keppel, 1982, pp. 158-159), then the Keppel technique should be chosen because of its relative power. This indicates that under the complete null hypothesis the recommended procedure would be to: 1) test the omnibus F first and if significant then, 2) test all simple effects without a penalty, then 3) for each significant simple effect, test the pairwise differences (simple comparisons) within the simple effect for significance with no penalty.

Note that the data were generated under conditions where the assumptions underlying the ANOVA were met. It has been shown that when there is a single factor design, these conclusions may not be upheld when the ANOVA assumptions are violated (e.g., Boik, 1981; Keppel, 1991). This qualification also applies to the results generated under the partial null hypothesis.

The Partial Null Hypothesis

When a more complex study is used (i.e. a factorial design), it may be more reasonable to assume that not all of the pairwise differences in means are zero. The partial null hypothesis represents the situation where null differences are embedded within the context of true interaction effects. Contrary to the results under the complete null hypothesis, the post hoc comparison techniques do not provide adequate protection against α_{FW} under the partial null hypothesis. The post hoc comparison techniques rely on the omnibus F and the simple effects tests to filter out false results. When there are true effects, the likelihood that the simple comparisons will be examined for significance increases because the probability of finding the omnibus F and simple effects significant will be more than .05.

The familywise error results under the partial hypothesis are complex, with the amount of protection afforded by the different techniques varying as a function of several factors: 1) type of design, 2) pattern of means, 3) effect size of the interaction, 4) effect size of the simple effect (null vs. true), 5) the number of null simple comparisons within each simple effect, and 6) the number of simple effects in which the null simple comparisons are embedded. The sixth factor appears to be the primary factor responsible for the results under the partial null hypothesis. The most important of these effects are discussed below.

Type of Design and Pattern

Overall (see Table 24), for 3 x 3 and 5 x 3 designs all techniques except Fisher serve to control familywise error under the partial null hypothesis when there is minimum to moderate variability across levels of a true simple effect (Pattern 1 and Pattern 2). If the means are at the extremes (maximum variability) for a 3 x 3, 5 x 3, or a 3 x 5 design, however the best techniques to control familywise error under the partial null hypothesis are DBON and Tukey Overall.

If one assumes that researchers seldom conduct a factorial design with more than three levels per variable and that the means are not at the extremes for true simple effects, then there should be little concern over choice of technique. Any control technique other than Fisher will work. When considering the power results, the best choice would be to use the Keppel approach – 1) test the interaction for significance and if significant, 2) test the simple effects for significance with no penalty, and 3) test the simple comparisons within each significant simple effect.

The only time that choosing a control technique becomes a problem is when the researcher exceeds three levels per variable or when the pattern of the means for the true simple effects represent maximum variability. When this occurs Tukey Overall is the only control technique that controls α_{FW} in all cases. Dual Bonferroni is a more powerful technique than Tukey Overall, but it fails to control familywise error in 3 x 5 designs when all 5 simple effects are true.

Effect Size of the Interaction

All techniques, except for Fisher and Keppel in 3 x 5 or 5 x 3 designs when the effect size of the interaction was small and where the means for true simple effects are at the extremes (Pattern 3), effectively control familywise error under the partial null hypothesis. The difference in power among the techniques was small when there was a small interaction effect size. It could be argued that almost any post hoc control technique could be used in this situation. However, given the difficulty of detecting true effects of this size, the most powerful technique should again be utilized—either Bonferroni or Modified Bonferroni.

When the effect size of the interaction is moderate (.25) or greater, the choice of control technique depends on the type of design and the pattern of the means across levels of the simple effect (see previous discussion). Based upon published literature, Cohen (1988) found that psychological research seldom results in effect sizes greater than .40, which he considers to represent a large effect size for psychological research. It could be argued, therefore, that the results regarding the .60 interaction effect size are outside the bounds of most psychological research. Even if this is the case, the results for moderate (.25) and large (.40) interaction effect sizes serve to point out that the choice of post hoc comparison control procedure is non-trivial.

Null vs. True Simple Effects

The pattern of the partial null familywise error results varied as a function of whether or not the null effects were embedded within the context of a true simple effect. All techniques except Fisher controlled familywise error under the partial null hypothesis

for 3 x 3 and 5 x 3 designs. The same was true when the effect size was small and the null simple comparisons occurred within the context of a null simple effect (see Table 20). For the 3 x 5 design, however, when the null simple comparisons were spread over three null simple effects only the simple effect penalty techniques (Bon, MB, MBB), dual penalty techniques (DBON, DMB) and Tukey Overall adequately controlled α_{FW}. Of the aforementioned techniques, Bonferroni, Dual Bonferroni, and Tukey Overall appear to the best for controlling familywise error. When power is also considered, Bonferroni appears to be the best choice among the three previously mentioned control techniques for the 3 x 5 design. To make things even easier, because there is no difference for the 3 x 3 and 5 x 3 designs, Bonferroni should also be used for these designs when there is no true simple effect.

The results are more complex when there are null simple comparisons embedded within the context of a true simple effect (Table 22). All techniques sufficiently controlled familywise error under the partial null hypothesis (except for Fisher) when there was a small interaction effect size. However, when the interaction effect size was moderate or larger, only the simple comparison penalty techniques (TROW, BROW), Dual Bonferroni (DBON) and Tukey Overall could control α_{FW} in all designs. One would probably choose either TROW, BROW, or DBON after considering the power results. The recommendation would be, therefore, to use one of these techniques when there is a true simple effect.

In summary, these results suggest that the approach used should depend on whether or not there is a true simple effect. Aside from using Tukey Overall on every occasion, the results indicate that when there is no true effect, Bonferroni should be used,

and when there is a true effect, either Tukey Row, Bonferroni Row, or Dual Bonferroni should be the techniques of choice. Looking back, this difference makes intuitive sense. When there is no true simple effect, paying a penalty at the level of simple effects (i.e., BON) leads to testing no simple comparisons within the null simple effect. Alternatively, when there is a true simple effect and the error rate penalty is not sufficient, all simple comparisons will be tested without any protection unless a simple comparisons penalty is applied (TROW, BROW, DBON).

From a pragmatic point of view, however, problems can arise when using this information as a guideline. The first problem is that it is impossible to determine in advance when there is a true or null simple effect. The second problem is that this indicates that one would have to change his or her analysis approach within any given study depending upon whether or not there is a presumed simple effect. Finally, if one adopted a single overall approach based on this information one would need to know in advance the relative proportion of null to true simple effects.

Overall – Number of Simple Effects in Which the Null Simple Comparisons are Located

When conducting a study, a researcher does not know the number of null and true simple effects. Knowing this, using the overall familywise error results is probably more prudent. The factor that most directly affects the magnitude of familywise error under the overall study of the partial null hypothesis appeared to be the number of simple effects that contain null simple comparisons. This makes intuitive sense because simple effects are independent of other simple effects, therefore null simple comparisons from different

simple effects are also independent. As previously stated in the results section, if effects are independent then they add more to familywise error than redundant effects.

The magnitude of familywise error was relatively low, in general, if the null simple comparisons were isolated within one simple effect. In this case, all techniques except Fisher worked to control familywise error. Even Keppel sufficiently controlled familywise error because the simple effects filter was enough to reduce α_{FW} from its low base rate, even without a penalty.

The magnitude of familywise error was high when the null simple comparisons were spread over three simple effects. In this case, the null comparisons are more likely to be independent, thereby increasing α_{FW} at a faster rate. When this occurs, it is necessary to pay a penalty at the level of simple effects to reduce familywise error to an acceptable level, with BON being the best choice.

When the null simple effects were spread across 5 simple effects, however, there were no techniques that effectively controlled familywise error other than TOVL, which is a very conservative test.

When viewing this information from a from a practical viewpoint it becomes apparent that the information is not very useful. How does one know, for example, if the null simple comparisons are isolated within one, three, or five simple effects? In order to use this information when choosing a post hoc statistical procedure one would need to give careful consideration to the pattern of results expected for all conditions before beginning the study. This is basically predicting the outcome of the study in advance. If this is the situation, then why wouldn't the researcher utilize planned comparisons and test only the comparisons of interest?

Why Not Use Tukey Overall?

Tukey Overall appears to be the best choice for controlling α_{FW} under the partial null hypothesis for all the designs investigated in this thesis. The ability of Tukey Overall to control familywise error was apparent, and it didn't matter whether or not there was a true simple effect or if the analysis was overall. The drawback to Tukey Overall is that it was the least powerful technique of all the techniques investigated. Most researchers agree that a willingness to find a true difference when one exists (power, the compliment of Type II error) versus finding a null simple effect should influence the choice of techniques for controlling familywise error. Some researchers argue that choosing a technique that balances the two types of errors is the best approach. However as Keppel (1991) states,

> In post-hoc data analysis, the type of question asked shifts from "Is *this* difference significant?" which characterizes planned comparisons, to "*Which* differences are significant?" The concern is with the whole set of treatments rather than with one condition with a particular combination of conditions . . . It is my opinion that post hoc comparisons should be subjected to a more stringent standard to guard against committing an unacceptably large number of Type I errors. (p.183)

If Keppel's position is adopted and one wants to simplify the procedure of selecting a control technique, then Tukey Overall should be the procedure of choice. By doing so, however, there is an enormous loss of power (22-42% for a large effect size). When using Tukey Overall a researcher should test the omnibus F for significance, then directly test the simple comparisons using the Tukey Overall penalty for the total number of means tested.

Type II Error and Power

The results of the current investigation indicate that studies conducted by psychologists are under powered (not adequately designed to detect true treatment effects at the level of simple comparisons for the designs studied) if a sample size of 8 and 15 represent typical sample sizes for psychological research. Some authors believe that it is appropriate to design a study so the researcher has an 80% chance (power = .80) of finding a true treatment effect. In the current study, a researcher would need an effect size of at least .65 to achieve acceptable power when using the Planned Comparison approach for a sample size of 8.

Type I error is inversely related to Type II error and directly related to power, which explains why the ordering of power for the 10 control techniques are predictable in terms of Type I error. The differences in power among the techniques were minimal (approximately 2-3 %) for small effect sizes because there was a such great likelihood of committing a Type II error (approximately .97-.98).

When the effect sizes were moderate to very large, testing only the simple effects for significance did not significantly affect the magnitude of Type II error or power (Fisher vs. Keppel). However, in some cases, one could affect Type II error and power if a simple effects penalty was paid instead of testing the simple effects as a contingency (Keppel vs. Bonferroni). Paying a penalty at the level of simple comparisons had an even larger effect on Type II error and power than paying a penalty at the level of simple effects (Keppel vs. Bonferroni, Keppel vs. Bonferroni Row and Bonferroni vs. Bonferroni Row). Dual Bonferroni (a dual error rate penalty) yielded lower Type II error rates than

Bonferroni (a simple effect penalty), but not larger Type II error penalties than a simple comparison penalty technique (Bonferroni Row). Clearly, Tukey Overall lead to the greatest Type II error and the least power.

Future Research

There are two possible lines of research that could be used in the future. The first follows evidence that suggests that the majority of problems linked to the post hoc control techniques occurred when the ANOVA contained an independent variable with five levels (i.e., a 3 x 5 or a 5 x 3 design) when controlling for familywise error under the partial null hypothesis. All techniques but Fisher adequately controlled for familywise error under the partial null hypothesis for the 3 x 3 design when the means were equally spaced (minimum/moderate variability) across levels of true simple effects. There is still the question about an independent variable with four levels. Would it yield the same results as a three or five level variable? A possible study, therefore, could be extended to present the line of research with 4 x 3 and 3 x 4 designs.

The second line of research could extend this research to three or more factors. When the number of factors in a factorial design are increased the number of conditions are increased resulting thereby increasing the likelihood of committing a Type I error. When there are three independent variables, Keppel recommends that the omnibus F is followed by analyses of simple interaction effects, then analysis of simple effects, and finally simple comparisons. There is an additional filter (analysis of simple effects) which may be effective in reducing the magnitude of Type I error under the partial null hypothsis, with this testing strategy. Additionally, it became important to determine at

what level an error rate penalty would be most effective, if one were applied? In comparison to Tukey overall, would this testing strategy be more probable to cause a reduction in familywise error rate under the partial null hypothesis?

Summary and Conclusion

Familywise error under the partial null hypothesis appears to be the major problem with post hoc analytic procedures. The control techniques do not give sufficient protection (using .05 as the familywise criterion for acceptability) when there are null simple comparisons embedded within the context of true interaction effects. The recommendation would be to use the Keppel approach for the 3 x 3 designs when the means for the true simple effects are not at the extremes. For any other situation tested in this study (3 x 5 design, 5 x 3 design, or a 3 x 3 design where the means are at the extremes), however, the only technique which consistently controlled familywise error was Tukey overall. The use of Tukey overall is not without a cost, however, as there is a large loss in power to detect significance.

This research confirms Keppel's conclusion that effects are so complex that it is difficult to reach a conclusion when a researcher exceeds three levels per variable. Therefore when planning a factorial design it is best to limit the number of levels to three per variable.

REFERENCES

Anthony, C., & Biers, D.W. (1995). [Comparison of methods for controlling familywise error in factorial designs]. Unpublished raw data.

Boik, R. (1981). A priori tests in repeated measures designs: Effects of nonsphericity. Psychological Bulletin, 86, 1084 – 1089.

Brake, G.L. (1994). *A comparison of methods for controlling familywise error: Post-hoc analyses for factorial designs*. Unpublished masters thesis, University of Dayton, Dayton, OH.

Carmer, S.G., & Swanson, M.R. (1973). An evaluation of ten pairwise multiple comparison procedures by Monte Carlo methods. Journal of the American Statistical Association, 68, 66-74.

Cohen, J. (1988) Statistical power analysis for the behavioral sciences (2^{nd} ed.). Hillsdale, NJ: Lawrence Erlbaum Associates.

Fisher, R.A. (1951). The design of experiments (6^{th} ed.). Edinburgh: Oliver& Boyd.

Hayter, A. (1986) The maximum familywise error rate of Fisher's least significant differences test. Journal of the American Statistical Association, 81, 1000 – 1004.

Jaccard. J., Becker, M.A., & Wood, G. (1984). Pairwise multiple comparison procedures: A review. Psychological Bulletin, 96, 589-596.

Keppel, G. (1982). Design and Analysis: A researcher's handbook (2nd ed.). Englewood Cliffs, NJ: Prentice Hall.

Keppel, G. (1991). Design and Analysis: A researcher's handbook (3rd ed.). Englewood Cliffs, NJ: Prentice Hall.

Keselman, H.J., Games, P.A., & Rogan, J.C. (1980). Type I and Type II errors in simultaneous and two-stage multiple comparison procedures. Psychological Bulletin, 88, 356-358.

Keselman, H.J., Keselman, J.C.., & Games, P.A. (1991). Maximum familywise Type I error rate: The least significant difference, Newman-Keuls, and other multiple comparison procedures. Psychological Bulletin, 110, 155-161.

Kirk, R.E. (1982). Experimental design: Procedures for the behavioral sciences (2nd ed.). Monterey, CA: Brooks/Cole.

Knuth, D. E. (1973). The Art of Computer Programming (2nd ed.). Reading, MA: Addison-Wesley.

Petrinovich, L.F., & Hardyck, C.D. (1969). Error rates for multiple comparison methods: Some evidence concerning the frequency of erroneous conclusions. Psychological Bulletin, 71, 43-54.

Ramsey, P.H. (1981). Power of univariate pairwise multiple comparison procedures. Psychological Bulletin, 90, 352-366.

Reising, J.D. (1993). *Alternative methods for controlling compounding error rate in post-hoc analysis of complex experimental designs: A Monte Carlo simulation of Type I and Type II errors.* Unpublished masters thesis, University of Dayton. Dayton, OH.

Rosnow, R.L., & Rosenthal, R. (1989). Definition and interpretation of interaction effects. Psychological Bulletin, 105, 143-146.

Ryan, T.A. (1980). Comment on "Protecting the overall rate of Type I errors for pairwise comparisons with an omnibus statistic." Psychological Bulletin, 88, 354-355.

Sheffé, H. (1953). A method for judging all contrasts in analysis of variance. Biometrika, 40, 87-104.

Stevens, J. (1986). Applied multivariate statistics for the social sciences. Hillsdale, NJ: Lawrence Erlbaum Associates.

Winer, B.J. (1972). Statistical principles in experimental design. New York: McGraw Hill.

Zwick, R., & Marascuillo, L.A. (1984). Selection of pairwise multiple comparison procedures for parametric and non-parametric analyses of variance models. Psychological Bulletin, 95, 148-155.